OLGYAY-FEKETE JUDIT[1], YANG ZIJIAN GYŐZŐ[2], ROBIN EDINA[3]

[1] ELTE BTK Nyelvtudományi Doktori Iskola, Fordítástudományi Doktori Program
E-mail: feketejudit@student.elte.hu
https://orcid.org/0009-0005-9696-5185

[2]HUN-REN Nyelvtudományi Kutatóközpont
E-mail: yang.zijian.gyozo@nytud.hun-ren.hu
https://orcid.org/0000-0001-9955-860X

[3]ELTE BTK Nyelvi Közvetítés Intézete, Fordító- és Tolmácsképző Tanszék
E-mail: robin.edina@btk.elte.hu
https://orcid.org/0000-0003-2025-4457

Olgyay-Fekete Judit – Yang Zijian Győző – Robin Edina: Gépi fordítás, utószerkesztés és lektorálás – humán és gépi kiértékelés
Machine translation, post-editing and revision – human and automated evaluation
Alkalmazott Nyelvtudomány, Különszám, 2024/3. szám, 135–151.
doi:http://dx.doi.org/10.18460/ANY.K.2024.3.008

Gépi fordítás, utószerkesztés és lektorálás: humán és gépi kiértékelés

Machine translation, post-editing, and revision – human and automated evaluation

In the wake of the technological turn in translation services, machine translation has become widespread. Companies, public administrations, and even professional translation agencies have integrated it into their workflow, as the use of AI speeds up the translation process enormously. Due to continuous technological advances, machine translation systems are also constantly improving, although not yet reaching the quality of human translation. However, the gap is narrowing and it is becoming increasingly difficult to determine the exact quality of machine translated texts using merely automated methods: quality evaluation algorithms based on reference translations and quality estimation. Automatic metrics offer fast and easy-to-perform analyses, while the human method provides detailed, qualitative data. The present pilot study was carried out within the joint research project of the Department of Translation and Interpreting of Eötvös Loránd University and the European Commission's Directorate-General for Translation (DGT). The aim of the Human-in-the-Translation-Loop (HITTL) research project is to investigate whether revision is necessary after post-editing machine pre-translated texts (Robin et al., 2023). In this paper, different automated quality evaluation methods – BLEU, chrF, TER, and neural quality estimation – were compared with the results of human analyses on domain-specific data. We also examined whether automated methods reflect the differences between machine translated, post-edited and revised texts. The results show that, at the document level, reference-based methods reflect quality variations well, in line with the human analyses, whereas quality estimation methods cannot do the same. Moreover, the results of the preliminary research clearly show an improvement in the quality of the target language text versions thanks to the quality assurance procedures: the weighted error rate shows a steady decrease as a result of the post-editing and proofreading. The difference in the quality of target text versions is lowest for the revised translations, i.e. the desired text quality is the result of the combination of post-editing and revision. This leads to the conclusion that revision of post-edited texts by a linguist other than the post-editor/translator is still necessary for optimal translation quality.

Keywords: machine translation, post-editing, revision, quality assessment, quality estimation

1. Bevezetés

A fordítóiparban lezajló technológiai fordulat (Jiménez-Crespo, 2020) új fordítási módszereket, minőségbiztosítási intézkedéseket és új szerepeket eredményezett a professzionális fordítási szolgáltatások területén (Eszenyi, 2023). A mesterséges intelligencián alapuló gépi fordítás széles körben elterjedt, immár nem csupán a laikus felhasználók hétköznapi kommunikációját elősegítő vagy a weboldalak automatikus fordítását szolgáló, könnyen elérhető alkalmazásnak számít, hanem beépült a professzionális nyelvi közvetítő eszköztárába is. A gépi fordítás idő- és költségtakarékos megoldásokat kínál, így nem meglepő, hogy a friss felmérések szerint fordításszolgáltatók és nemzetközi cégek is alkalmazzák (ELIS, 2024).

A digitális technológia, elsősorban a mesterséges intelligencia fejlődésével a gépi fordítórendszerek minősége is folyamatosan fejlődik. Az ember és a gép által fordított szövegek közötti minőségi különbség egyre csökken, de a gépi fordítás még nem éri el az emberi fordítás színvonalát és megbízhatóságát. Ez különösen igaz a műfordításra és a doménspecifikus szövegekre. Ahogyan csökken azonban a szakadék, úgy a kiértékelő módszereknek is egyre nehezebb a dolguk, rendkívül finom különbségek között kell meghatározniuk a minőségbeli eltéréseket.

A gépi fordítómotorok által készített fordítások kiértékelése több szempontból is fontos a szakma és a tudomány számára egyaránt (Yang, 2024). A különféle motorok különböző minőséget produkálnak, ezért fontos megállapítani, hogy egy bizonyos szövegtípus esetében melyik fordítómotor kimenete lehet megfelelő az adott célra. Nem elhanyagolható ugyanis, hogy egy nagyobb fordítási projektben milyen minőségű gépi fordítást kell utószerkesztenie a nyelvi szakembernek. Egy rossz minőségű szöveg akár több munkát és erőfeszítést is jelenthet a fordítónak, mint ha önállóan készítené el a fordítást, tehát a technológia nem éri el a célját: a hatékonyság növelését és a költségek csökkentését. Az értékelések eredménye a fordítómotorok készítői számára is fontos adatokkal szolgál, hiszen tudniuk kell, mennyire hatékony a rendszerük, milyen újításokra, fejlesztésekre van szükség. A munkájuk javításához fontos kiindulópontot adhat, ha a nyelvészek rá tudnak világítani a gépi fordítás erősségeire és gyengeségeire.

A gép által fordított szövegek minőségének értékelésekor megkülönböztetünk automatikus és manuális, azaz gépi és humán módszereket. A nyelvtechnológián alapuló automatikus értékelés gyors és könnyen elvégezhető elemzéseket kínál, a humán módszer pedig részletes, kvalitatív adatokkal szolgál. A gépi fordítás minőségbiztosításának fontosságát jelzi, hogy megindult a törekvés az értékelés szabványosítására. Külön konzorcium jött létre, hogy egységes minőségértékelési szempontrendszert és eszközkészletet biztosítson a szakma és a kutatók számára (Lommel, 2018). A nemrégiben megjelent ISO 5060:2024 minőségbiztosítási szabvány a szakfordítások humán értékelésének és lektorálásának egységesítését segíti elő, hangsúlyozva egyúttal a jelentőségét is.

A jelen pilotkutatás az ELTE BTK Fordító- és Tolmácsképző Tanszékének és az Európai Bizottság Fordítási Főigazgatóságának (DGT) közös projektjeként,

illetve egyik részeként valósult meg. A Human-in-the-Translation-Loop (HITTL) kutatócsoport azt vizsgálja, hogy van-e szükség lektorálásra a gépi fordítómotor által készített fordítás utószerkesztése után (Robin et al., 2023). A kutatási projekt egyik alkérdéseként azt vizsgáltuk meg, hogy a referenciafordításon alapuló gépi minőségértékelő algoritmusok és a neurális minőségbecslési módszerek vajon képesek-e kimutatni az utószerkesztett és a lektorált szövegváltozatok közötti minőségi változást, illetve korrelálnak-e az alkalmazott hibatipológián alapuló humán minőségértékelés eredményeivel.

2. Automatikus minőségértékelés és minőségbecslés

A gépi fordítás kiértékeléséhez a mai napig használatosak a referenciafordításon alapuló minőségértékelő metrikák. Ilyen a BLEU (Papineni et al., 2002), a chrF (Popović, 2015), a TER (Snover et al., 2006) vagy a METEOR (Banerjee & Lavie, 2005), illetve e metrikák azon változatai, amelyek nem a gépi fordítás és a referencia közötti különbséget mérik, hanem a gépi fordítás és az utószerkesztett szöveg közötti eltéréseket. Ezek a *Human-targeted* előtaggal ellátott változatok: HBLEU, HTER, HMETEOR. A legnépszerűbb metrikák mellett továbbiakat is használhatnak különböző, specifikusabb nyelvekre vagy feladatokra: NIST (Doddington, 2002), ROUGE (Lin, 2004), LEPOR (Han et al., 2012), RIBES (Isozaki et al., 2010). Az ilyen metrikák érvényességének egyik sarkalatos pontja az összehasonlítás alapjául szolgáló referenciafordítások eredete és minősége.

A gépi fordítás kiértékelésének másik módszere a referenciafordítás nélküli kiértékelés, más néven minőségbecslés. A minőségbecslő modell felépítése során gépi tanuló algoritmussal minőségi mutatók vagy jegyek alapján tanítanak be egy modellt emberi kiértékelésekre (Yang et al., 2016; Specia et al., 2013). Az utóbbi években egyértelműen a neurális módszereken alapuló minőségbecslő módszerek érik el a legjobb eredményeket. A minőségbecslő modell két különböző nyelvű szöveget hasonlít össze, ezért a többnyelvű előtanított nyelvmodelleken alapuló módszerek váltak népszerűvé (Rei et al., 2022; Tao et al., 2022). Másik kutatási irány, amikor a két különböző nyelv számára két külön enkódert (*dual encoder*) tanítanak be (Heo et al., 2021). A többnyelvű modellek tovább kombinálhatók többfeladatos tanulási architektúrákkal (*multitask learning architectures*) (Lim et al., 2021; Geng et al., 2022), vagy saját, manuálisan előállított jegyekkel lehet tovább bővíteni a modell tudását (Wang et al., 2021; Zerva et al., 2021).

A gépi metrikáknak – a humán kiértékeléshez hasonlóan – vannak előnyeik és hátrányaik is. Előnyük, hogy gyorsan el lehet őket készíteni (gyakorlatilag néhány gombnyomás segítségével), míg a humán kiértékelés (jelentős mennyiségű) idő- és energiaráfordítást igényel. A gépi metrikák továbbá bármikor rendelkezésre állnak, és objektív adatokkal szolgálnak. Hátrányuk azonban, hogy nem tudnak árnyalt értékelést szolgáltatni; nem tudják megállapítani a hibák típusát, illetve nem tudják a hibákat súlyozni sem (egy félrefordítás például súlyosabb hiba, mint egy elütés). Egyes metrikák pedig nem veszik figyelembe például a szórendet, vagy büntetőpontot adnak, ha a fordítás rövidebb a referenciaszövegnél (BLEU).

Éppen ezért a könnyen alkalmazható gépi metrikák mellett nem elhanyagolható a humán kiértékelés szerepe sem a minőségbiztosítás terén.

3. Humán kiértékelés

Korábbi kutatások bizonyították, hogy a statisztikai módszereken alapuló gépi fordítórendszerek idejében a referenciafordításon alapuló metrikák csak kevéssé korreláltak a manuális elemzésekkel (Banerjee & Lavie, 2005; Laki, 2015). Az utóbbi években a gépi fordítórendszerek minőségének javulásával az értékelési szempontok is egyre részletesebbek lettek. A kutatók hosszú távú célja az, hogy az emberi és a gépi fordítás kiértékelését egységesíteni, szabványosítani lehessen. Ennek érdekében került bevezetésre a gépi fordítás minőségértékelésébe az MQM Core (*Multidimensional Quality Metrics*) hibatipológiai szempontrendszer (lásd Freitag et al., 2021; Yang, 2024), amely szó-, kifejezés- és mondatszintű elemzést is lehetővé tesz. Ez a hibatipológia szolgált alapul az ISO 5060:2024 nemzetközi minőségbiztosítási szabvány értékelési szempontrendszerének kialakításához is.

A humán kiértékelés mindig pontosabb, mint a gépi kiértékelés, ám a hátránya, hogy időigényes és drága. Szükségességét azonban a *human-in-the-loop* modell[1] is alátámasztja, amely alapján a mesterséges intelligencia fejlesztéséhez nagyban hozzá tud járulni a humán nyelvi szakember munkája az adatok annotálásánál, az algoritmus tanításánál, majd az algoritmus tesztelésénél (Mosqueira-Rey, 2023).

Way (2018) idézi a gépi fordítás humán kiértékelésének típusait, amelyeket Humphreys és munkatársai (1991) tettek közzé. Ezek a típusok a következők:

- Tipológiai értékelés (*typological evaluation*): megmutatja, mely fordítási jelenséget tudja kezelni az adott gépi fordítórendszer.
- Deklaratív értékelés (*declarative evaluation*): azt mutatja meg, hogy egy gépi fordítómotor hogyan teljesít a fordítás különböző dimenziói mentén.
- Operatív értékelés (*operational evaluation*): meghatározza, hogy egy gépi fordítómotor mint a fordítási folyamat része mennyire lesz hatékony a költségek tekintetében.

Meglátásunk szerint a fordítóipar jelenleg abba az irányba halad, hogy a nyelvi szolgáltatók számára a harmadik típus a leglényegesebb. Talán ez az irány nem éppen a legkedvezőbb, legpozitívabb a nyelvtudomány szempontjából, az ipari körülményeket azonban nem hagyhatjuk figyelmen kívül, különösen akkor nem, ha olyan szakembereket szeretnénk képezni, akik a fordítópiacra kikerülve meg tudják állni a helyüket (lásd még Eszenyi et al., 2023). Az emberi kiértékelésnek az időigényesség és a költségesség mellett másik hátránya a szubjektivitás, éppen ezért elengedhetetlen, hogy a megbízhatóság érdekében a hibatipológiai elemzést több nyelvi szakember végezze. Popović (2018) szerint a humán kiértékelés akkor okozza a legnagyobb nehézséget, ha sokféle hibatípust kell megkülönböztetni,

[1] https://www.telusinternational.com/glossary/human-in-the-loop

hiszen ilyenkor sokkal nagyobb a tévedés lehetősége, és az annotátorok között ilyenkor a legkisebb mértékű az egyetértés. Popović cikkében azon álláspontját is ismerteti, amely szerint az automatikus és a humán kiértékelés kombinációjával lehet a legpontosabb és legátfogóbb értékelést adni a gépi fordítás minőségéről.

Falkedal és King (1990) már korábban megfogalmazták ezen véleményüket, ugyanakkor pesszimistábbak a megvalósítás vonatkozásában. Szerintük is az a legjobb megoldás, ha több olyan szakember végzi az értékelést, aki rendelkezik nyelvészeti ismerettel, és van tapasztalata a nyelvtechnológia terén. Ám hozzá is teszik, hogy ez a két feltétel csak a legritkább esetben valósul meg a gyakorlatban. A helyzet ma is változatlan: ritka az együttműködés a nyelvészek, a gyakorló fordítók és a nyelvtechnológusok között. A fordítóiparban végbemenő digitális forradalom azonban szükségessé teszi a szakterületek közötti együttműködést: a mesterséges intelligencia átalakítja a fordítási folyamatokat, a szakma megoldást keres a minőségbiztosítás kérdéseire, a nyelvi kérdésekre pedig a fordításkutató nyelvészek keresnek választ. Kutatásunkkal remélhetőleg szorosabbra fűzhető a különböző szakterületek közötti kapcsolat.

4. A kutatás bemutatása

A jelen pilotkutatás a HITTL kutatócsoport projektjéhez (Robin et al., 2023) kapcsolódva arra a kérdésre keresi a választ, hogy a referenciafordításon alapuló automatikus minőségértékelő algoritmusok, valamint a neurális minőségbecslési módszerek mennyire követik az utószerkesztett és a lektorált szövegváltozatok közötti különbségeket, illetve mennyire korrelálnak az alkalmazott hibatipológián alapuló humán értékelés eredményeivel. Az automatikus gépi kiértékeléseken és manuális hibatipológiai szövegelemzéseken alapuló kismintás előzetes kutatás elsődleges célja a kiválasztott módszerek tesztelése volt.

4.1. A korpusz

Kutatásunkhoz három angol nyelvű sajtóközleményt használtunk fel, amelyeket az Európai Bizottság Fordítási Főigazgatósága (DGT) bocsátott a kutatócsoport rendelkezésére, így lehetővé vált, hogy valós fordítási projekteket vonjunk be a vizsgálatokba. A DGT nyelvi szakemberei integrált fordítási környezetben, CAT-eszközök segítségével végzik a fordítást (Ábrányi, 2015; Robin et al., 2023), de a sajtóközlemények esetében kevés fordítómemória áll a rendelkezésükre, ezért a fordítók nagyrészt a gépi fordítómotor fordításaira hagyatkoznak. A kutatási korpusz a DGT neurális fordítómotorjával (eTranslation)[2] fordított szövegeket, valamint azok utószerkesztett és lektorált változatát tartalmazza. Az automatikus és gépi elemzési módszereink tesztelését szolgáló pilotkutatás három európai uniós sajtóközlemény magyar nyelvű változatainak vizsgálatára korlátozódott, a szövegek jellemzői az alábbi 1. táblázatban láthatók. A vizsgálatokból kizártuk a

[2] eTranslation – European Commission https://commission.europa.eu/resources-partners/etranslation_en

fordítómemóriából származó szegmenseket, és csak a gépi fordítás bevonásával készült szövegrészeket vizsgáltuk.

A különböző szövegváltozatok minőségének kiértékelését kétféle elemzési módszerrel végeztük el. Az egyik a manuális, humán hibatipológiai elemzés volt, a másik a különböző metrikákat alkalmazó gépi minőségbecslés és -értékelés. Az automatikus minőségértékelő metrikákhoz a fordítások hibatipológiai elemzése eredményeként szuperlektorált változatát használtuk fel referenciaszövegként.

1. táblázat. A korpusz főbb szövegjellemzői

Dokumentum (doc_id)	Szövegváltozatok	Elemzett szegmensek	Szó	Átlagos szószám / szegmens (átlag/medián)
1.	Forrás	38	936	24,63 / 24,0
	Gépi fordítás		870	22,90 / 22,5
	Utószerkesztett		874	23,00 / 22,5
	Lektorált		875	23,03 / 22,5
	Szuperlektorált (referencia)		873	22,97 / 21,5
5.	Forrás	17	342	20,12 / 20,0
	Gépi fordítás		292	17,18 / 17,0
	Utószerkesztett		289	17,00 / 16,0
	Lektorált		302	17,77 / 17,0
	Szuperlektorált (referencia)		304	17,88 / 18,0
9.	Forrás	51	924	18.12 / 16,0
	Gépi fordítás		803	15,75 / 15,0
	Utószerkesztett		839	16,45 / 15,0
	Lektorált		843	16,53 / 15,0
	Szuperlektorált (referencia)		858	16,82 / 15,0

4.2. Humán minőségértékelés

A szövegek humán kiértékelését a DGT által is alkalmazott (Drugan et al., 2018) MQM Core hibatipológia[3] segítségével hajtottuk végre, amely a humán és a gépi fordítás elemzésére is alkalmazható. A tipológiát a kutatás céljai és az elemzések eredményei alapján adaptáltuk. Az eredeti tipológiában szereplő fő kategóriák közül négyet használtunk a vizsgálatban, figyelmen kívül hagyva a szerkesztésre vonatkozó pontokat, mivel a szövegeket projektcsomagként, szerkesztés nélküli változatban kaptuk meg. A kutatásban használt hibakategóriák az alábbiak voltak:

[3] https://themqm.org/the-mqm-typology/

- Pontosság (*accuracy*)
  - Félrefordítás (*mistranslation*)
  - Túlfordítás (*over-translation*)
  - Alulfordítás (*under-translation*)
  - Betoldás (*addition*)
  - Kihagyás (*ommission*)
  - Szükségtelen fordítás (*unnecessary translation*)
  - Nemfordítás (*non-translation*)
  - Zavaros fordítás (*garbled translation*)
- Terminológia (*terminology*)
  - Inkonzisztens terminus (szövegen belül) (*inconsistent term within text*)
  - Inkonzisztens terminus (forrással) (*inconsistent term with resource*)
  - Helytelen terminus (*wrong term*)
  - Hiányos terminus (*incomplete term*)
  - Szükségtelen terminus (*unnecessary term*)
  - Hiányzó terminus (*missing term*)
- Nyelvi norma (*linguistic norm*)
  - Nyelvtan (*grammar*)
  - Központozás (*punctuation*)
  - Helyesírás (*spelling*)
- Nyelvhasználat (*style*)
  - Szervezeti nyelvhasználat (*organisational style*)
  - Inkonzisztencia (forrással) (*inconsistent style with external reference*)
  - Ügyetlen nyelvhasználat (*awkward style*)
  - Nem idiomatikus nyelvhasználat (*unidiomatic style*)
  - Inkonzisztencia (szövegen belül) (*inconsistent style within text*)
  - Regiszter (*register*)

A hibatipológiai elemzéseket a kutatócsoport tagjainak részvételével, többszörös kódolással végeztük a megbízhatóság és a szubjektivitás kizárása érdekében.

A kategorizálás során a hibák súlyozására is sor került, hiszen a minőséget nem ugyanolyan mértékben veszélyezteti a központozást és a szakmai terminológiát érintő hiba. Az MQM Core tipológia a hibák súlyozását illetően négy kategóriát különböztet meg (kis, nagy, kritikus, semleges), az ISO 5060:2024 nemzetközi szabvány pedig hármat (kritikus, nagy, kis), azonban a DGT által alkalmazott értékelési rendszert követve a kategorizálás egyszerűsítése érdekében csupán kis és nagy hibákat azonosítottunk. Nagynak minősültek az értelemzavaró hibák, például a félrefordítás és a helytelen terminológia, míg kicsinek az értelmezést nem veszélyeztető központozási vagy helyesírási hibák számítottak. Az összesítés során a kis hibák egyes, a nagy hibák pedig ötös szorzóval kerültek számításba.

4.3. Gépi minőségbecslés és -értékelés

Kutatásunkban a referenciafordítással történő automatikus kiértékelő metrikák közül az alábbiakat alkalmaztuk:

- **BLEU** (Papineni et al., 2002): a BLEU (*BiLingual Evaluation Understudy*) az egyik legnépszerűbb kiértékelő módszer. A BLEU pontosságot számol, azt vizsgálja, hogy a gép által lefordított mondatokban lévő szavak és kifejezések mennyire pontosan illeszkednek a referenciafordításhoz. Az algoritmus az n-grammok (*n* szomszédos szimbólum) alapján számolt értékek súlyozott átlagát adja. A vizsgálathoz a SacreBLEU (Post, 2018) implementációt alkalmaztuk.
- **chrF** (Popović, 2015): a chrF algoritmus karakter n-gramm F-mértéket számol. Ez a módszer a ragozó nyelveknél különösen előnyös, hiszen ha két szó csak toldalékban különbözik, akkor azokat a BLEU eltérő szóalakokként kezeli, és nem talál közöttük egyezést, míg a chrF sokkal több egyezést talál a karakterek szintjén. A jelen kutatáshoz az eredeti implementációt[4] használtuk, vagyis nem módosítottunk a paramétereken.
- **TER** (Snover et al., 2006): a TER (*Translation Edit Rate / Translation Error Rate*) fordítási hibaarányt számol a gépi fordítás és az emberi referenciafordítás között az alapján, hogy mennyi szóalapú eltérés (beszúrás, törlés, eltolás vagy helyettesítés) azonosítható, majd az azonosított eltérések számát elosztja a referenciafordítások átlagos mondathosszával. Kutatásunkhoz a SacreBLEU-féle implementációt (Post, 2018) alkalmaztuk.
- **xCOMET-XXL** (Guerreiro et al., 2023): az Unbabel[5] terméke; a 10,7 milliárd paraméteres XLM-R XXL (Goyal et al., 2021) modellt finomhangolták gépi fordítások kiértékelésére. A tanításhoz a WMT22-r[6] által kiadott *Direct Assessments* és MQM korpuszt használták. Leírásuk alapján támogatja a magyar nyelvet is. A kiértékeléshez a forrásnyelvi, a gépileg lefordított és a referenciaszöveget kell megadni a modellnek.

A fentiekben ismertetett automatikus vizsgálatokhoz a humán kiértékelés alapján létrehozott szuperlektorált szövegváltozatokat használtuk.

A kutatásban alkalmazandó, referenciafordítás nélküli minőségbecslő modell kiválasztásának legfontosabb szempontja az volt, hogy a modell ismerje az angol és a magyar nyelvet is, ezért a választást két modellre szűkítettük le:

- **QuEst**: angol–magyar minőségbecslő modell. Yang és Laki (2023) kutatásuk során egy XLMRoBERTa (Conneau et al., 2020) modellt finomhangoltak az angol–magyar HuQ (Yang et al., 2016) korpuszon.

[4] https://github.com/m-popovic/chrF
[5] https://unbabel.com
[6] https://www.statmt.org/wmt22

- **CometKiwi-da-xxl** (Rei et al., 2023): az Unbabel[7] terméke; a 10,7 milliárd paraméteres XLM-R XXL (Goyal et al., 2021) modellt finomhangolták minőségbecslés feladatára. Bemenetként a forrásnyelvi és a gépileg lefordított szöveg szolgál az elemzéshez.

A kiértékelés során kimértünk még két többnyelvű modellt is az Unbabeltől, amelyek támogatják a magyar nyelvet. Mindkét modell XXL méretű, több mint 44 GB és több mint 10 milliárd paraméteres. A használatukhoz egy darab nagy teljesítményű, 80 GB-os NVIDIA A100 GPU szerverre volt szükség.

5. Eredmények

A kiválasztott három EU-s sajtóközlemény különböző – eTranslation rendszerrel előfordított, utószerkesztett, lektorált és szuperlektorált – változatait kiértékeltük a gépi minőségértékelő és -becslő metrikákkal, valamint humán hibatipológiai elemzést is végeztünk az MQM Core előre meghatározott kategóriái alapján. A gépi és az emberi minőségértékelések összehasonlításhoz a Pearson-korrelációt alkalmaztuk. (A korrelációs együttható 1 és -1 közötti értékeket vehet fel.)

A humán kiértékelés eredményei a 2. táblázatban láthatók, a kis és nagy hibák számát összesítve, és meglehetősen változatos képet mutatnak. A gépi fordításnál minden hibakategóriában magasabb értékek láthatók, mint az utószerkesztett és a lektorált szövegváltozatoknál, világosan jelezve a minőségbiztosítási eljárások szükségességét a fordítási folyamatban: a gépi fordítás önmagában még nem tud megfelelő minőségű célnyelvi szöveget előállítani. Az utószerkesztett fordítások esetében a hibák száma csökken, de még magasabb, mint a lektorált szövegekben. A hibatipológiai elemzés adatai alapján az is jól követhető, hogy a különböző szövegekben eltérő gyakorisággal fordulnak elő az egyes hibatípusok.

[7] https://unbabel.com

2. táblázat. A humán kiértékelés eredményei

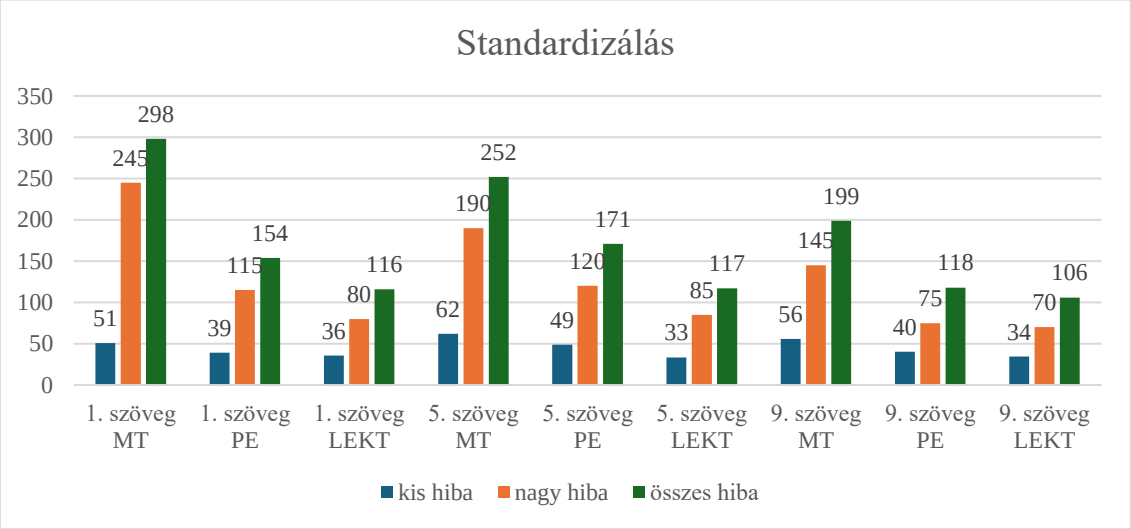
Doc_id	Hibatípus	Gépi fordítás	Utószerkesztett	Lektorált
1. (935 szó)	Pontosság	14	11	8
	Terminológia	22	10	8
	Nyelvi norma	33	19	15
	Nyelvhasználat	19	16	16
	Összesen	88	56	47
5. (342 szó)	Pontosság	6	5	3
	Terminológia	2	1	1
	Nyelvi norma	5	5	2
	Nyelvhasználat	16	10	9
	Összesen	29	21	15
9. (952 szó)	Pontosság	10	7	7
	Terminológia	5	6	6
	Nyelvi norma	19	11	9
	Nyelvhasználat	34	23	19
	Összesen	68	47	41

Az 1. gépileg előfordított szövegben a nyelvi normát sértő hibák szerepeltek a legnagyobb számban, legkevesebb a fordítás pontossága, vagyis az ekvivalencia terén mutatkozott. Az 5. szövegben a nyelvhasználatot érintő hibák voltak döntő többségben, illetve a terminológián belül fordult elő a legkevesebb probléma a gépi fordításban. A 9. szöveg ugyanazt a képet mutatja, mint az 5. gépi fordítás: a legtöbb hiba a nyelvhasználaton belül, a legkevesebb pedig a terminológia terén fordult elő. Az eredmények valószínűleg azzal magyarázhatók, hogy a betanított neurális gépi fordítómotorok szószinten könnyen azonosítják és feleltetik meg a terminusokat, és a fejlesztések eredményeként a nyelvek közötti átváltás sem okoz olyan gondokat, mint a korábbi modelleknél. A mondatszintű problémák, az idiomatikus megfogalmazás, a megfelelő regiszter és a feldolgozhatóság viszont továbbra is gyakori problémaként jelenik meg a gépi fordításban.

A vizsgált szövegek minőségének pontos megítéléséhez azonban szükséges a hibák súlyozása is (lásd 4.2. alfejezet). Az elemzések során tehát a kis hibákat egyes, a nagy hibákat ötös szorzóval láttuk el. Továbbá a különböző terjedelmű

vizsgált szövegek összehasonlíthatósága érdekében a kapott értékeket 1000 szóra standardizáltuk. Az így kapott súlyozott számadatokat az 1. ábra mutatja be.

1. ábra. Standardizált eredmények



Az oszlopdiagramon megfigyelhető a vizsgált szövegváltozatok minőségének javulása a minőségbiztosítási eljárások nyomán: a hibaértékek alacsonyabbak az utószerkesztett és a lektorált változatokban minden szöveg esetében. Ez a javulás elsősorban a nagy hibák számában bekövetkező csökkenéssel magyarázható. Jól látható a különbség az egyes szövegek minőségében is: az 1. előfordított szöveg (MT) súlyozott hibaértéke a legmagasabb, a 9. szövegé a legkisebb – az eltérés pedig ebben az esetben is a nagy hibák számértékének eltérésével magyarázható. Érdemes megfigyelni azt is az adatok szórása alapján, hogy az utószerkesztés és a lektorálás nyomán csökken a minőségbeli különbség a szövegváltozatok között: gépi fordítások (49,5), utószerkesztett szövegek (27,1), lekorált szövegek (6,1).

A 3. táblázatban látható a gépi és a humán kiértékelések összehasonlítása. A referenciaalapú automatikus metrikák magas korrelációval (0,9<) követték az egyes szövegváltozatok minőségi változását, amelyet a humán hibatipológiai elemzés eredményei megmutattak. A neurális minőségbecslő rendszerek azonban nem tudták egyértelműen felismerni a minőségjavulást (az 5. szöveg kivételével). Az angol–magyar minőségbecslő modellt ugyanis a HuQ korpuszon tanították be, amely nem tartalmaz neurális gépi fordításokat, ez a modell valószínűleg inkább a statisztikai modellekkel lefordított szövegek hibáit ismeri fel. A 9. szöveg esetében a gépi minőségbecslés és a humán elemzés ugyan magas korrelációt ért el, de negatívat, vagyis a metrikák fordítva érzékelték a minőségváltozást.

3. táblázat. Gépi kiértékelő metrikák eredményei

Doc_id	Metrika	Gépi fordítás	Utószerkesztett	Lektorált	Referencia	Pearson-korreláció
1	MQM	71,379	84,554	88,457	100	-
	BLEU	70,874	81,052	85,294	100	0,9890
	chrF	84,964	91,147	93,345	100	0,9980
	TER	23,368	15,578	11,111	0	0,9890
	xCOMET-XXL	0,909	0,948	0,955	0,980	0,9955
	QuEst	5,037	5,031	5,042	5,040	0,3939
	CometKiwi-da-xxl	0,912	0,932	0,922	0,921	0,4055
5	MQM	76,370	83,045	88,411	100	-
	BLEU	63,297	74,236	81,479	100	0,9996
	chrF	81,366	86,924	91,123	100	0,9998
	TER	27,961	19,079	12,171	0	0,9979
	xCOMET-XXL	0,961	0,980	0,985	0,992	0,9102
	QuEst	5,076	5,095	5,0967	5,097	0,7567
	CometKiwi-da-xxl	0,952	0,967	0,975	0,978	0,8990
9	MQM	80,697	88,439	89,443	100	-
	BLEU	36,657	51,014	52,606	100	0,8636
	chrF	51,408	59,283	59,990	100	0,9439
	TER	82,051	72,494	70.746	0	0,9273
	xCOMET-XXL	0,662	0,668	0,668	7,862	0,8711
	QuEst	5,063	5,0589	5,0594	5,054	-0,9944
	CometKiwi-da-xxl	0,8154	0,8910	0,8931	0,6206	-0,6966

A 4. táblázat a szegmensalapú kiértékelések eredményeit mutatja. Itt az látható, hogy a gépi elemzések szegmensalapon vegyesen tudtak magas korrelációt elérni az emberi kiértékeléssel. A referenciaalapú metrikák az 1. szövegnél viszonylag magas (0,7<) korrelációt mutatnak a gépi fordítás és az utószerkesztett szöveg esetében is, míg az 5. szövegnél csak a gépi fordításról mondható el ugyanez. A 9. szövegnél rendkívül alacsony korrelációt értek el a metrikák. Az angol–magyar minőségbecslő modell egyetlen esetet leszámítva mindenhol alacsony korrelációt talált. Az xCOMET viszont magas korrelációkat ért el az 1. és az 5. szövegnél, de a 9. szöveg esetében ugyancsak alacsony értékeket mutatott. A számok alapján a 9. szöveg minőségének mérése nehéznek bizonyult mindegyik modell számára.

A szegmensalapú kiértékelésből az látszik egyértelműen, hogy a minőségbecslő modellek nem tudták követni az emberi kiértékeléseket.

4. táblázat. Szegmensalapú Pearson-korreláció eredmények

Doc_id	Metrika	Gépi fordítás	Utószerkesztett	Lektorált
1	BLEU	0,743	0,769	0,687
	chrF	0,857	0,899	0,367
	TER	0,768	0,728	0,641
	xCOMET-XXL	0,689	0,770	-0,044
	QuEst	0,124	-0,118	0,050
	Cometkiwi-da-xxl	0,096	-0,059	-0,013
5	BLEU	0,733	0,313	0,247
	chrF	0,657	0,307	0,154
	TER	0,772	0,283	0,234
	xCOMET-XXL	0,376	0,657	0,319
	QuEst	-0,078	0,343	0,492
	Cometkiwi-da-xxl	-0,099	0,552	0,005
9	BLEU	0,337	0,166	0,370
	chrF	0,029	0,027	0,282
	TER	0,049	0,107	0,387
	xCOMET-XXL	-0,088	-0,041	0,265
	QuEst	-0,066	-0,087	-0,132
	Cometkiwi-da-xxl	-0,304	-0,216	-0,227

6. Összegzés

A jelen kutatásban különböző gépi minőségértékelő módszereket vizsgáltunk a humán elemzéssel összehasonlítva abból a szempontból, hogy mennyire követik a gépileg előfordított, az utószerkesztett és a lektorált szövegváltozatok közötti minőségjavulásokat. A pilotkutatásnak nem volt célja azonban a hibatipológiai elemzés kvalitatív adatainak részletes bemutatása – erre egy későbbi tanulmány keretében kerülhet sor. Vizsgálatunkban a humán hibatipológiai elemzés mellett referenciaalapú és referencia nélküli automatikus minőségértékelő, illetve -becslő módszereket is alkalmaztunk. Eredményeink azt mutatják, hogy a referenciaalapú értékelő metrikák a dokumentum szintjén tudják követni a humán elemzéssel is

azonosított minőségjavulást, a minőségbecslő modell azonban nem. Ez annak tudható be, hogy az általunk alkalmazott minőségbecslő modell tanítóanyaga, a HuQ Korpusz nem tartalmaz neurális fordítómotorokkal készült fordításokat.

Az előzetes kutatás eredményei világosan megmutatják a szövegváltozatok minőségének javulását a minőségbiztosítási eljárásoknak köszönhetően: a hibák súlyozott értéke folyamatos csökkenést mutat az utószerkesztés és a lektorálás eredményeképpen. A szövegváltozatok minősége közötti különbség a lektorált szövegek esetében a legalacsonyabb; a kívánt szövegminőség az utószerkesztés és a lektorálás együttes eredményeként jött létre. Ebből pedig azt a következtetést vonhatjuk le, hogy szükség van az utószerkesztett szövegek lektorálására.

A jelen pilotkutatás csak az Európai Bizottság Fordítási Főigazgatósága (DGT) által fejlesztett eTranslation neurális gépi fordítórendszer vizsgálatára fókuszált, a különböző szövegváltozatok minőségi javulását elemezve. A jövőben érdemes lehet más, általános és doménspecifikus fordítómotorok, valamint a generatív mesterséges intelligencián alapuló rendszerek bevonásával folytatni a kutatást. Azt is fontos megvizsgálni a későbbiekben, hogy a minőségértékelő metrikák képesek-e követni az egyes szövegek közötti, humán hibatipológiai elemzéssel azonosított minőségi különbségeket, valamint kimutatni az utószerkesztés és a lektorálás sajátosságait (Szlávik 2022). Emellett több fordított szöveg elemzésére van szükség, mivel a különböző általánosítható tendenciák megállapítása csupán nagyobb korpuszok elemzésével lehetséges. Ehhez a DGT által biztosított további gépileg előfordított szövegeket tervezzük megvizsgálni a jövőben.

Irodalom

Ábrányi Henrietta (2015). Fordítási környezetek. In Horváth Ildikó (szerk.), *Modern fordító és tolmács*. (147–160). Budapest: ELTE Eötvös Kiadó.

Banerjee, S., Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: Goldstein, J., Lavie, A., Lin, C.Y., Voss, C. (Eds.), *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization* (65–72). Ann Arbor, Michigan: Association for Computational Linguistics. Letöltés: https://aclanthology.org/W05-0909

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (8440–8451). doi: 10.18653/v1/2020.acl-main.747.

Doddington, G. (2002). Automatic evaluation of machine translation quality using ngram co-occurrence statistics. In *Proceedings of the Second International Conference on Human Language Technology Research* (138–145). HLT '02, San Francisco: Morgan Kaufmann Publishers.

Drugan, J., Strandvik, I., Vuorinen, E. (2018). Translation Quality, Quality Management and Agency: Principles and Practice in the European Union Institutions. In Moorkens, J., Castilho, S., Gaspari, F., Doherty, S. (Eds.), *Translation Quality Assessment. Machine Translation: Technologies and Applications*, Vol 1. Springer, Cham. doi: https://doi.org/10.1007/978-3-319-91241-7_3

ELIS Survey (2024) *European Language Indutsty Survey.* Letöltés: https://elis-survey.org/

Eszenyi Réka (2023). *Humán Fordító és gépi fordítás 8 leckében. Változások a 21. századi nyelvi közvetítő szerepében.* Budapest: ELTE Eötvös Kiadó. doi: 10.21862/Transl.HuXMach.2023.8

Eszenyi, R. Bednárová-Gibová, K., Robin, E. (2023). Artificial intelligence, machine translation & cyborg translators: a clash of utopian and dystopian visions. *Orbis Linguarum, 21*(2), 103–112. doi: https://doi.org/10.37708/ezs.swu.bg.v21i2.13

Falkedal, K., King, M. (1990). Using test suites in evaluation of machine translation systems. In Falkedal, K., King, M. (Eds.), *COLING 1990 Volume 2: Papers presented to the 13th International Conference on Computational Linguistics*, Letöltés: https://aclanthology.org/C90-2037

Freitag, M., Foster, G., Grangier, D., Ratnakar, V., Tan, Q., Macherey, W. (2021). Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. In *Transactions of the Association for Computational Linguistics*, *9*, 1460–1474. doi: https://doi.org/10.1162/tacl_a_00437

Geng, X., Zhang, Y., **Huang, S., Tao, S., Yang, H., Chen, J.** (2022). NJUNLP's participation for the WMT2022 quality estimation shared task. In Barrault, L., Bojar, O., Bougares, F. Chatterjee, R., Costa-jussà, M.R., Federmann, C., Fishel, M., Fraser, A., Freitag, M., Graham, Y., Grundkiewicz, R., Guzman, P., Haddow, B., Huck, M., Kocmi, T., Koehn, P., Martins, A., Morishita, M., Monz, C., Nagata, M., Nakazawa, T., Negri, M., Névéol, A., Neves, M., Popel, M., Turchi, M., Yepes, A. J., Zampieri, M. (Eds.), *Proceedings of the Seventh Conference on Machine Translation (WMT).* (615–620). Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Goyal, N., Du, J., Ott, M., **Anantharaman, G., Conneau, A.** (2021). Larger-scale transformers for multilingual masked language modeling. In Bansal, T., Calixto, I., Camburu, O.M., Kassner, N., Rogers, A., Saphra, N., Shwartz, V., Vulić, I. (Eds.), *Proceedings of the 6th Workshop on Representation Learning for NLP* (RepL4NLP-2021) (29–33). Bangkok: Association for Computational Linguistics. Letöltés: https://aclanthology.org/2021.repl4nlp-1.4

Guerreiro, N. M., Rei, R., van Stigt, D,. **Coheur, L., Colombo, P., Martins, A.F.T.** (2023). *xcomet: Transparent machine translation evaluation through finegrained error detection*. doi: https://doi.org/10.48550/arXiv.2310.10482

Han A. L. F., Wong D. F., Chao, L. S. (2012). LEPOR: A robust evaluation metric for machine translation with augmented factors . In *Proceedings of COLING 2012* (441–450). Mumbai, India: The COLING 2012 Organizing Committee. 441–450.

Heo, D., Jung, B., Lee, J.H., Lee, W. (2021). Quality estimation using dual encoders with transfer learning. In Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussa, M. R., Federmann, C., Fishel, M., Fraser, A., Freitag, M. Graham, Y., Grundkiewicz, R., Guzman, P., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Kocmi, T., Martins, A., Morishita, M., Monz, C. (Eds.), *Proceedings of the Sixth Conference on Machine Translation* (920–927). Online: Association for Computational Linguistics. doi: 10.1162/tacl_a_00437

Humphreys, L., Jäschke, M., Balkan, L., Way, A., Meyer, S. (1991). *Operational evaluation of MT, draft research proposal.* Working papers in language processing 22, University of Essex.

Isozaki, H., Hirao, T., Duh, K., Sudoh, K., Tsukada, H. (2010). Automatic evaluation of translation quality for distant language pairs. In Li, H., Màrquez, L. (Eds.), *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (944–952). Cambridge, MA, USA: Association for Computational Linguistics, Letöltés: https://aclanthology.org/D10-1092

Jiménez-Crespo, M. A. (2020). The "technological turn" in translation studies. Are we there yet? A transversal cross-disciplinary approach. *Translation Spaces, 9*, 314–341. doi: https://doi.org/10.1075/ts.19012.jim

Laki László János (2015). *Statisztikai gépi fordítás módszerének alkalmazása egy- és többnyelvű nyelvtechnológiai problémák hatékony megoldására*. Doktori értekezés, Budapest: Pázmány Péter Katolikus Egyetem.

Lim, S., **Kim, H., Kim, H.** (2021). Papago's submission for the WMT21 quality estimation shared task. In Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussa, M. R., Federmann, C., Fishel, M., Fraser, A., Freitag, M. Graham, Y., Grundkiewicz, R., Guzman, P., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Kocmi, T., Martins, A., Morishita, M., Monz, C. (Eds.), *Proceedings of the Sixth Conference on Machine Translation* (935–940). Online: Association for Computational Linguistics. Letöltés: https://aclanthology.org/2021.wmt-1.98

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In Lin, C.-Y. (Ed.), *Text Summarization Branches Out* (74–81). Barcelona, Spain Association for Computational Linguistics. Letöltés: https://aclanthology.org/W04-1013

Lommel, A. (2018). Metrics for Translation Quality Assessment: A Case for Standardising Error Typologies. In Moorkens, J., Castilho, S., Gaspari, F., Doherty, S. (Eds.), *Translation Quality Assessment. Machine Translation: Technologies and Applications*, Vol 1. (109–127). Springer, Cham. doi: https://doi.org/10.1007/978-3-319-91241-7_6

Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D. (2023). Human-in-the-loop machine learning: a state of the art. *Artif Intell Rev, 56*, 3005–3054. doi: https://doi.org/10.1007/s10462-022-10246-w

Papineni, K., Roukos, S., Ward, T., Zhu, W. J. (2002). Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (311–318). Stroudsburg, PA, USA: ACL '02, Association for Computational Linguistics. Letöltés: http://dx.doi.org/10.3115/1073083.1073135

Popović, M. (2015). chrF: character n-gram F-score for automatic MT evaluation. In Bojar, O., Chatterjee, R., Federmann, C., Haddow, B., Hokamp, C., Huck, M., Logacheva, V., Pecina, P. (Eds.), *Proceedings of the Tenth Workshop on Statistical Machine Translation* (392–395). Lisbon, Portugal: Association for Computational Linguistics, Letöltés: https://aclanthology.org/W15-3049

Popović, M. (2018). Error Classification and Analysis for Machine Translation Quality Assessment. In Castilho, S., Doherty, S., Gaspari, F., Moorkens, J. (Eds.), *Translation Quality Assessment. Machine Translation: Technologies and Applications*, Vol 1. (1–30). Springer, Cham. doi: 10.1007/978-3-319-91241-7_7

Post, M. (2018). A call for clarity in reporting BLEU scores. In Bojar, O., Chatterjee, R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Monz, C., Negri, M., Névéol, A., Neves, M., Post, M., Specia, L., Turchi, M., Verspoor, K. (Eds.), *Proceedings of the Third Conference on Machine Translation: Research Papers* (186–191). Belgium, Brussels: Association for Computational Linguistics. Letöltés: https://www.aclweb.org/anthology/W18-6319

Rei, R., Guerreiro, N. M., Pombal, J., van Stigt, D., Treviso, M., Coheur, L., de Souza, J. G. C., Martins, A.F.T. (2023). *Scaling up cometkiwi: Unbabel-ist 2023 submission for the quality estimation shared task*. In *Proceedings of the Eighth Conference on Machine Translation* (841–848). Singapore. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.73

Rei, R., Treviso, M., Guerreiro, N.M., Zerva, C., Farinha, A. C., Maroti, C., de Souza, J.G., Glushkova, T., Alves, D., Coheur, L., Lavie, A., Martins, A.F.T., (2022). CometKiwi: IST-unbabel submission for the quality estimation shared task. In Barrault, L., Bojar, O., Bougares, F. Chatterjee, R., Costa-jussà, M.R., Federmann, C., Fishel, M., Fraser, A., Freitag, M., Graham, Y., Grundkiewicz, R., Guzman, P., Haddow, B., Huck, M., Kocmi, T., Koehn, P., Martins, A., Morishita, M., Monz, C., Nagata, M., Nakazawa, T., Negri, M., Névéol, A., Neves, M., Popel, M., Turchi, M., Yepes, A. J., Zampieri, M. (Eds.), *Proceedings of the Seventh Conference on Machine Translation (WMT)* (634–645). Abu Dhabi: Association for Computational Linguistics.

Robin Edina, Eszenyi Réka, Kóbor Márta, Seidl-Péch Olívia (2023). Human in the Translation Loop: az ELTE FTT és a DGT kutatási projektje. Elhangzott: TransELTE 2023. Budapest: Eötvös Loránd Tudományegyetem. (2023. március 2.)

Snover, M., Dorr, B., , Schwartz, R., Makhoul, J., Micciulla, L. (2006). A study of translation edit rate with targeted human annotation. In Dorr, B., Makhoul, J., Micciulla, L., Schwartz, R., Snover, M. (Eds.), *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers* (223–231). Cambridge, Massachusetts, USA: Association for Machine Translation in the Americas, Letöltés: https://aclanthology.org/2006.amta-papers.25

Specia, L., **Shah, K., de Souza, J. G., Cohn, T.** (2013). QuEst – a translation quality estimation framework. In Butt, M., Hussain, S. (Eds.), *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (79–84). Sofia, Bulgaria: Association for Computational Linguistics. Letöltés: https://aclanthology.org/P13-4014

Szlávik Szilárd. (2022). A gépi fordításhoz kötődő alapvető terminusok, definíciók és a közöttük lévő ellentmondások. *Fordítástudomány*, *24*(1), 87–103.

Tao, S., Chang, S., Miaomiao, M., Yang, H., Geng, X., Huang, S., Zhang, M. Guo, J., Wang, M., Li, Y. (2022). CrossQE: HW-TSC 2022 submission for the quality estimation shared task. In Barrault, L., Bojar, O., Bougares, F. Chatterjee, R., Costa-jussà, M.R., Federmann, C., Fishel, M., Fraser, A., Freitag, M., Graham, Y., Grundkiewicz, R., Guzman, P., Haddow, B., Huck, M., Kocmi, T., Koehn, P., Martins, A., Morishita, M., Monz, C., Nagata, M., Nakazawa, T., Negri, M., Névéol, A., Neves, M., Popel, M., Turchi, M., Yepes, A. J., Zampieri, M. (Eds.), *Proceedings of the Seventh Conference on Machine Translation (WMT)* (646–652). Abu Dhabi: Association for Computational Linguistics

Wang, J., Wang, K., Chen, B., Zhao, Y., **Luo, W., Zhang, Y.** (2021). QEMind: Alibaba's submission to the WMT21 quality estimation shared task. In Barrault, L., Bojar, O. Bougares, F., Chatterjee, R., Costa-jussa, M. R., Federmann, C., Fishel, M., Fraser, A., Freitag, M., Graham, Y., Grundkiewicz, R., Guzman, P., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Kocmi, T., Martins, A., Morishita, M., Monz, C. (Eds.), *Proceedings of the Sixth Conference on Machine Translation* (948–954). Association for Computational Linguistics. Letöltés: https://aclanthology.org/2021.wmt-1.100

Way, A. (2018). Quality Expectations of Machine Translation. In Moorkens, J., Castilho, S., Gaspari, F., Doherty, S. (Eds.), *Translation Quality Assessment. Machine Translation: Technologies and Applications*, Vol 1. Springer, Cham. doi: https://doi.org/10.1007/978-3-319-91241-7_8

Yang Zijian Győző (2024). A gépi fordítás és kiértékelésének módszerei a fordítástudományban. In Klaudy K., Robin E., Seidl-Péch O. (szerk.), *Bevezetés a fordítás és a tolmácsolás kutatásmódszertanába II. Speciális rész* (253–274). Budapest: ELTE FTT – MANYE Fordítástudományi Szakosztály. doi: 10.21862/kutamodszertan2/14

Yang, Z. G., **Laki, J. L.** (2023). *Enhancing machine translation with quality estimation and reinforcement learning*. Annales Mathematicae et Informaticae, Accepted manuscript.

Yang, Z. G., Laki, J. L., Siklósi, B. (2016). HuQ: An English-Hungarian corpus for quality estimation. In Bojar, O., Burchardt, A., Dugast, C., Federico, M., van Genabith, J., Haddow, B., Hajic, J., Harris, K., Koehn, P., Negri, M., Popel, M., Rehm, G., Specia, L., Turchi, M., Uszkoreit, H. (Eds.), *Proceedings of the LREC 2016 Workshop - Translation Evaluation: From Fragmented Tools and Data Sets to an Integrated Ecosystem* (43–49.) Portorož.

Zerva, C., **van Stigt, D., Rei, R., Farinha, A. C., Ramos, P., de Souza, J. G. C., Glushkova, T., Vera, M., Kepler, F., Martins, A. F. T.** (2021). IST-unbabel 2021 submission for the quality estimation shared task. In Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussa, M. R., Federmann, C., Fishel, M., Fraser, A., Freitag, M. Graham, Y., Grundkiewicz, R., Guzman, P., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Kocmi, T., Martins, A., Morishita, M., Monz, C. (Eds.), *Proceedings of the Sixth Conference on Machine Translation* (961–972). Online: Association for Computational Linguistics, Letöltés: https://aclanthology.org/volumes/2021.wmt-1/