Language analysis required to improve AI resources (Initially: Submission 12)

Marija Liudvika Drazdauskiene liudvika@drazdauskiene.lt

Responding to the leitmotif of this Conference, the question of whether linguists have something to offer to AI has been taken to focus in this paper. The source study was based on linguistic analysis. Three texts, *A Preface* from the book, *The Truths We Hold*, by Kamala Harris and two excerpts from a modern novel, *A Spark of Light*, by Jodi Picoult were taken from Google translate, from English into Lithuanian, studied, edited and assessed. In accord with the known truth, the excerpt from discursive prose was translated better than those from fiction, but translation inaccuracies were typical in the three texts of two different styles. Mechanical choices were obvious in equivalent selection as the inaccurate meaning of time because of the obvious missing options of lexico-syntactical and lexico-morphological variants in tense choices in computer corpora. Reasoning within corpus-based MT functioning as an example-based model, the following inaccuracies in TL appeared typical:

1 Lexical mistakes in both texts, which suggested mechanical choices:

"She's gone?" "I'm sorry." = „Jos nėra?" „Atsiprašau." / „Gaila."

„You just ripped my scrubs off." = "Tu ką tik nuplėšei mano šveitiklius." / „Tu ką tik nuplėšei mano kombinezoną."

„Hold still, Doctor." = „Laikykitės, daktare." / „Pagulėkit ramiai, daktare."

„There," she said. „That's more like it.' = "Ten."… "Tai labiau panašu." / „Va." ... „Jau beveik gerai." (JP)

… Alexander came up to me with big tears welling in his eyes = A. priėjo prie manęs su didelėmis ašaromis akyse / A. Priėjo prie manęs su pilnomis ašarų akimis.

Come here, little man. What's wrong? = Ateik čia, mažasis žmogau. Kas negero? / Kas nutiko? (KH)

2 Limited or missing resources of synonyms, which result in false equivalence:

She rolled the fabric from Dr Ward's pants leg... = Ji apvyniojo audinį nuo... kelnių klešnės... / Ji nuvyniojo drobę nuo...

... as she finished taping the tourniquet. = … baigdama klijuoti turniketą. / ...baigdama tvirtinti turniketą. (JP)

… when I heard one of Doug's signature groans = ... kai išgirdau vieną iš D. firminių *dejavimų / kai išgirdau vieną įprastinių D. dejavimų.*

... while poles hadn't yet closed in California = kol Kalifornijos balsavimai dar nebuvo užsidarę / kol Kalifornijos balsavimai dar nebuvo pasibaigę. (KH)

3 Differences in contextual meaning of synonyms as questionable semantic accuracy:

His voice broke on the last word. = Po paskutinio žodžio jo balsas nutrūko. / Sulig paskutiniu žodžiu jo balsas užlūžo. (JP)

None of us had fully processed what was happening = Nė vienas iš mūsų nebuvo iki galo apdorojęs to, kas vyksta / Nė vienas... nebuvo suvokęs to, kas vyko.

... we had a new mountain to take = turėjome pakilti į naują kalną / turėjome pakilti į naują aukštumą. (KH)

4 Differences in the formal tense meaning and lexico-morphological units of time:

He raised a brow. = Jis pakėlė antakį. / Jis kilstelėjo antakį.

Then she looked at her wrist. = Ji pažvelgė į savo riešą. / Ji žvilgtelėjo į riešą. (JP)

Something was happening – something bad. = Kažkas vyksta – kažkas blogo / Kažkas vyko – kažkas blogo.

... many of us cried tears of joy = daugelis iš mūsų verkė iš džiaugsmo ašarų / daugelis iš mūsų verkė džiaugsmo ašaromis. (KH)

Minding the negative highlights, linguists have something to offer to improve MT in developing electronic lexical corpora: increasing the corpus of lexical equivalents; detailing the meaning of synonyms and working on their technical programming; doing analysis and working out detailed description of the lexico-grammar of time expression in synthetic languages given that SL is English. Structural and lexical interdependence of the expresion of time researched in English (Carter, McCarthy, 2007) is missing in minor synthetic languages. Advancement in MT and AI poses new questions and demands new reseach, (Progress in machine translation, Elsevier, 2022), not mentioned here. Two ways may be foreseen contributing to AI: to improve and enlarge the existing electronic corpora for corpus-based MT and to create model text-based corpora which AI empowered super computers can process on the level of unit equivalence.

Key words: machine translation: semantic deviations: errors, false equivalence, questionable semantic accuracy; the semic analysis of meaning: lexical synonyms, lexico-grammatical synonyms; editing

(321 words)