ktu faculty of social sciences, arts and humanities 1922

43/2023
Research Journal
Studies about Languages
pp. 77–92
ISSN 1648-2824 (print)
ISSN 2029-7203 (online)
DOI 10.5755/j01.sal.1.43.35154

TRANSLATION / VERTIMAS

How much Romanian does *Google Translate* know? A corpus-informed genre-specific error analysis of English-into-Romanian translations

Received 09/2023 | Accepted 11/2023

HOW TO CITE: Pungă, L., Manda, I., & Chitez, M. (2023). How much Romanian does Google Translate know? A corpus-informed genre-specific error analysis of English-into-Romanian translations. *Studies about Languages / Kalbų studijos*, 43, 77–92. https://doi.org/10.5755/j01.sal.1.43.35154

How much Romanian does *Google Translate* know? A corpus-informed genre-specific error analysis of English-into-Romanian translations

Kaip gerai *Google Translate* moka rumunų kalbą? Tekstynu grįsta skirtingų žanrų tekstų vertimo iš anglų kalbos į rumunų kalbą klaidų analizė

LOREDANA PUNGĂ, West University of Timișoara, Romania

IONELA MANDA, West University of Timișoara, Romania

MĂDĂLINA CHITEZ, West University of Timișoara, Romania

Abstract To compensate, even if on a small scale, for the scarcity of investigations of English-into-Romanian machine translation from a corpus-based genre-specific perspective, this case study concerns the translation between these languages, by Google Translate, of texts belonging to two different, but closely-related genres – everyday and newspaper/ news releases language. In particular, it aims to offer a view of the translation errors in the Romanian target texts and, implicitly, of translation quality. To meet this aim, a translation error analysis is performed, starting from Keshavarz's (1999) very general model of error analysis, and a linguistic error profile is created for each of the two genres taken into consideration. The errors identified are discussed and illustrated with small-scale corpus examples. Since translation errors, affecting translation quality, are the direct consequence of the capabilities of the Google translation platform, the findings of this paper may be relevant for the developers of this platform. They may get a clearer picture of its strengths and limitations and suggest ways of improving it so that it can ultimately provide higher quality translations when working with the English – Romanian pair of languages, with the particular text genres looked at here. It may also contribute to raising translators' attention to the areas that are potentially problematic in these contexts, in the post-editing stage.

KEYWORDS: machine translation, Google Translate, translation errors, translation quality, Romanian.

Introduction

Since 2006, Google Translate has been making its contribution to breaking down language barriers and building bridges of communication around the globe. In 2008, Romanian was featured in this machine translation platform alongside other languages such as Bulgarian, Croatian, Czech, Danish, Finnish, Hindi, Norwegian, Polish, and Swedish. In 2022, Google Translate added 24 new (under-resourced) languages, reaching a coverage of 133 languages worldwide.

Google Translate started off as a statistical machine translation platform (SMT). Ten years later, in 2016, the owner company announced that its translation service would switch to Google Neural Machine Translation (GNMT), which means it then became able to take into account broader linguistic contexts, advancing from the translation of isolated words and phrases to the translation of entire clauses, no matter their length. Nowadays, this multilingual machine translation platform is able to provide the translation of multiple forms of text and media instantly, thus considerably reducing the amount of time and effort invested in this process by human translators. Though technological progress is evident in the case of Google Translate, what remains to be discussed is whether progress in the quality of the translations it provides has followed track.

With this in mind, our aim here is to offer a profile of errors identified in the translations, from English into Romanian, of texts belonging to two different, but closely-related genres – everyday and newspaper/ news releases language and, subsequently, to point at Google translation quality in the specific contexts considered.

Theoretical Background

In the large amount of current research on the topic of machine translation quality assessment, it has been pointed out, like in a recent Google Research report (Bapna et al., 2022) for example, that scores calculated automatically for the quality of machine-translation, such as BLEU, CHRF or the newly suggested loose or strict RTTLANGIDCHRF, are often not fully reliable by themselves and many focus on how well the automatic translation system functions rather than on the quality of the end product. Moreover, they cannot be calculated at all in the absence of reference translations to take into consideration. Many advocate the opinion that they fail to provide enough insight for error analysis, so that "any decision on translation quality ultimately cannot be made with automatic metrics [...] alone" (Bapna et al., 2022, p. 23), and a combined human – automated assessment is recommended to obtain more reliable results concerning the quality of the end product itself (Rivera-Trogueros, 2021; Chatzikoumi, 2020; Way, 2018). The need for additional human assessment (ideally, by native speakers of the target language) in terms of target text adequacy and fluency, accuracy, acceptability and readability is, thus, of paramount importance to make a final decision on the quality of machine-translated texts.

The analysis in this article stems from this very conviction that human assessment of machine translation, no matter how time- and energy-consuming in a world driven by speed and efficiency, is a necessary endeavour in deciding whether an adequate level of equivalence between the source and the target texts (from purely lexical and grammatical to stylistic and pragmatic) has been obtained.

In her survey of machine translation assessment methods, Chatzikoumi (2020) suggests that there are two broad categories of assessment methods that human judges resort to: methods based on directly-expressed judgement (DEJ-based evaluation methods) and methods that are not based on such judgments (non-DEJ-based evaluation methods).

As their name indicates, the former presuppose that judges directly express their opinion on translation quality, usually either by comparing the source text to the target text (bilingual assessment) or by comparing the target text to some reference text in the target language (monolingual assessment). The main aspects taken into consideration are accuracy and fluency, assessed on the basis of point-scales, of ranking (ideally no more than three) machine translation systems, comparing two machine translation systems or providing either general or more specific judgments of the translation.

In non-DEJ-based evaluation methods, human assessors express their judgements only indirectly. They may do this by using semi-automated metrics (known also as "human in the loop evaluation"); "by performing tasks which require the comprehension of a machine-translated text; or by classifying, analysing and correcting MT

outputs" (Chatzikoumi, 2020, p. 13). Semi-automated metrics are quite similar to automated ones, but they presuppose the involvement of human assessors (annotators) as well. Task-based non-DEJ evaluation relies on looking at the efficacy of the translated text when it comes to performing tasks like "asking people to detect the most relevant information in a text; asking people to answer questions on the text's content (Sanders et al. 2011); and gap filling, that is restoring keywords in reference translations (Ageeva et al., 2015)" (Chatzikoumi, 2020, p. 13). The classification, analysis and correction of MT output very often takes the form of error classification and analysis, be this done based on metrics (like the Multidimensional Quality Metrics – MQM) or following classifications of error typologies.

In this article, the approach we take is a combined DEJ-based and non-DEJ-based error evaluation, as the core of our analysis is represented by error classification and analysis, but judgments concerning the accuracy and fluency of the target texts are also made.

Research Methodology

Corpus. Sub-corpora

This study is based on a descriptive qualitative analysis that we have carried on a corpus of translations from Romanian into English, with the source texts selected, in the spring of 2021, so as to cover two genre-based categories: everyday communication and newspaper articles/ news agencies releases. The sub-corpus in the former category consists of 1,154 words, in the form of 103 sentences of various lengths and levels of (mostly lexical) complexity: idiomatic expressions, phrasal verbs, and proverbs frequently used in English, many of which lack direct counterparts in Romanian. The selection of these structures was based on publicly available lists, sourced from online repositories. Subsequently, context was either identified or created, based on the contextual appropriateness of the expressions.

The one in the latter category is a 6,641-word sub-corpus selected from the most recent news reports available at the time of the study, focusing on a range of topics so as to tackle as many fields of activity as possible: NP1 addresses racial issues, with a focus on anti-Asian racist incidents; NP2 narrates the US Capitol riot in January 2021, dealing with the events in a chronological, cause-and-effect, story-like manner; NP3 describes the Covid situation in Brazil, emphasizing the country's serious problems in managing the pandemic; NP4 reports on ecology and pollution in China; NP5 deals with gender issues as reflected in the use of language in LGBTQ communities and NP6 elaborates on the news of the cargo ship that blocked the Suez Canal in the spring of 2021, with an analysis of the event's disastrous economic consequences. The news reports were chosen so that they should ensure the inclusion of a broad spectrum of terminology and writing styles, thereby presenting a pretty wide range of challenges from a translation perspective.

Data analysis

The English source texts were translated using the digital translation tool provided by Google, Google Translate, the most popular worldwide free of charge machine translation platform (Rivera-Trigueros, 2021), with more than one billion users at the moment (Wise, 2023). The target texts in Romanian were analysed against Keshavarz's (1999) very general classification of translation errors, into four categories: (a) orthographic errors, (b) phonological errors, (c) lexico-semantic errors and (d) morphological-syntactic errors. Due to the nature of the corpus investigated here, phonological errors could not have been taken into consideration. The choice of this very broad classification out of many (eg. Secară, 2006; Snover et al., 2009; Koponen, 2010; De la Cruz-Cabanillas & Tejedo-Martínez, 2016; Popović, 2018; Dumitran, 2021, etc) is motivated by the fact that it offered us the possibility of further dividing the categories ourselves, with specific reference to our corpus and thus, to a specific language pair, in the two genre-based linguistic contexts. While Keshavarz (1999) clear-cuts the error categories, based on our findings, we indicate that some target language stretches contain errors that are simultaneously of various types and may, therefore, be included in more than one category (as can be seen in Tables 1 and 2).

The corpus-based analysis was carried out by employing the #LancsBox (Lancaster University corpus toolbox) software package (Brezina et al., 2021). #LancsBox is a freely-available new-generation corpus analysis tool, which offers a wide variety of built-in functions, making text annotation, formatting and encoding texts quite fast

and very efficient. This user-friendly software also offers essential tools in corpus linguistics, such as the generation of concordance lines (which it can sort, filter and randomise), visualisation of collocations and colligations, frequency lists of words, phrases, various morphological categories and complex linguistic structures, frequency of n-grams types, lemmas and POS categories, etc. (McEnery & Hardie, 2012), allowing for easy information retrieval and text exploration.

In the context of the present study, it has been used as an initial corpus scrutiny tool. We began the analysis by identifying the words most frequently used in our sub-corpora which are prone to translation errors. Then, a comparison between contextual uses of these frequent linguistics items was made. The comparative analysis involved looking at the corpus of translated versions against the corpus of original texts, encompassing both lexical and contextual aspects.

After the initial corpus-based frequency analysis and comparative corpus analysis, a manual analysis of the Google Translate generated target texts has been performed. Examples were extracted to illustrate the translation errors identified, with the final aim of getting a general opinion on the quality of the translation we have analysed. We did not aim at a quantitative or rank-based analysis of these errors.

Results and Discussion

Table 1 below includes the errors in the translation from English into Romanian, by the Google digital tool, of the everyday language sub-corpora.

Table 1 Errors in the Google translation of everyday language samples (English into Romanian)

Error type	ST (English)	TT (Romanian)
lexico-semantic	piece of cake	*bucată de tort*
	eye and ear candy	*bomboane pentru ochi și urechi*
	He got all bent out of shape over nothing.	*S-a aplecat cu totul din formă peste nimic.*
	This is not something anybody can wrap their head around until they have no choice but to get on board.	*Acest lucru nu este ceva ce oricine își poate înfășura capul până când nu are de ales decât să urce la bord.*
	We found some problems with the house that set the renovations back two weeks.	*Am găsit câteva probleme cu casa care au făcut renovările în urmă cu două săptămâni.*
	It's raining cats and dogs outside.	*Afară plouă pisici și câini.*
	the last straw	*ultimul pai*
	A bird in the hand is worth two in the bush, always remember that.	*O pasăre în mână valorează două în tufiș, amintiți-vă întotdeauna asta.*
	My mother walks out of the room when my father brings up sports.	*Mama iese din cameră când tatăl meu face sport.*
	would be off his rocker	*ar fi în afara rockerului său*
	sleeping dogs lie	*câinii adormiți să mintă*
	nail on the head	*unghia de cap*
	The woman broke down when the police told her that her son had died.	*Femeia s-a defectat când poliția i-a spus că fiul ei a murit.*
	I accidentally ran over your bicycle in the driveway.	*Ți-am alergat accidental peste bicicleta pe alee.*
	The police set up the car thief by using a hidden camera.	*Poliția a pus la cale hoțul de mașini folosind o cameră ascunsă.*
	Sorry, you're not fat, you're horizontally challenged.	*Ne pare rău, nu ești grasă, ești provocată orizontal.*

Error type	ST (English)	TT (Romanian)
morphological-syntactic + lexical	was under the weather	*va fi sub vreme*
	Well, according to Charlie, you and this Mrs Sparrow are like two peas in a pod.	*Ei bine, conform lui Charlie, tu și această doamnă Sparrow sunteți ca doi mazăre într-o păstăi.*
	Everyone seems to be trying to butter up the new boss hoping to become her favorite.	*Toată lumea pare să încerce să-l ungă pe noul șef sperând să devină preferatul ei.*
orthographic + lexico-semantic	You might cut her some slack.	*S-ar putea să-i tăieți* [tăiați] puțin.*
	chocolate bar	*batonă* [baton] de ciocolată*
	I have to hand in my essay by Friday.	*Trebuie să-mi predez* [predau] eseul până vineri.*

From a lexico-semantic perspective, one of the tasks the Google Translate platform seems to struggle with is rendering the idiomatic meaning of words and phrases in the source text into the target text, as it often fails to grasp their non-literal meaning. Consequently, it rather translates word by word in these instances. This results in a literal translation, divergent from the actual meaning of the source text phrases.

Ex 1 piece of cake
bucată de tort (lit.)

Ex 2 was under the weather
va fi sub vreme (lit.)

Ex 3 eye and ear candy
bomboane pentru ochi și urechi (lit.)

Ex 4 He got all bent out of shape over nothing.
S-a aplecat cu totul din formă peste nimic.
(He bent over completely from his shape over nothing.)

Ex 5 This is not something anybody can wrap their head around until they have no choice but to get on board.
Acest lucru nu este ceva ce oricine își poate înfășura capul până când nu are de ales decât să urce la bord.
(This thing is not something that anybody can wrap their heads (lit.) until they have no choice but to get on board.)

Ex 6 We found some problems with the house that set the renovations back two weeks.
Am găsit câteva probleme cu casa care au făcut renovările în urmă cu două săptămâni.
(We found some problems with the house that made the renovations two weeks ago.)

Ex 7 You might cut her some slack.
S-ar putea să-i tăieți puțin. (You may cut her a little bit.)

Example 7 is an interesting case of translation error – it is hard to decide whether the translation has been taken only half-way towards the target version or whether it has been on the wrong track from the very beginning. If one admits the former, what is missing is the clarification on what should be reduced. i.e. restrictions. A full, acceptable translation would then have been *S-ar putea/ Ați putea să-i tăiați puțin din restricții (It would be possible to / You may cut some of the restrictions)*, which takes over, in a non-idiomatic form, the idiomatic meaning of the original. One spelling mistake is present in this example, too – *tăieți* is the misspelt form of *tăiați*.

A case of combined lexical and morphological-syntactic error occurs in Example 2, where, besides the literal translation of the idiomatic "was under the weather", the tense of its verb, the past, is replaced by the future in Romanian.

Nevertheless, quite surprisingly, Google Translate does not behave similarly in all cases of idiomaticity. There are idioms which it translates correctly, offering indirect, non-literal equivalents for them (it remains to be investigated whether this is in some way connected to the use frequency of the idioms in question). Such is the case of the idiom "it's raining cats and dogs".

Ex 8 It's raining cats and dogs...
Plouă cu găleata... (approx. it pours rain as if from a bucket)

Still, if the translation unit is larger than the idiomatic phrase itself, even by a single word (be it a content or a form word), correct translation becomes problematic. This is what happens, for example, when the determiner "outside" is added to "it's raining cats and dogs", the word for word alternative is offered for the whole unit, resulting into a nonsensical, though grammatically correct, sequence in Romanian: *afară plouă pisici și câini*.

This is not an isolated case. Google Translate is able to provide the appropriate Romanian equivalent for the idiom "last straw" – *ultima picătură*, but once the definite article is added – "the last straw", the platform provides a literal translation – *ultimul pai*. Similarly, it correctly translates the proverb "A bird in the hand is worth two in the bush" into Romanian, but once this proverb is subordinated to a main clause (which is, indeed, unusually but not incorrectly situated after the secondary clause and separated by it by a comma in English), the translation platform fails to preserve its meaning and provides a fully literal alternative: "A bird in the hand is worth two in the bush, always remember that" – *O pasăre în mână valorează două în tufiș, amintiți-vă* întotdeauna *asta*.

Like sayings and proverbs, the greatest majority of English idiomatic comparisons have a set, non-literal equivalent in Romanian. "Like two peas in a pod", for instance, translates as *(seamănă) ca două picături de apă* ((they look alike) like two drops of water), a wording that is missed by the digital translation platform we have used in this study. The translation solution is not only incorrect from a lexical perspective, but from a grammatical one as well. Lexically, apart from not considering the idiomatic meaning of *two peas in a pod*, *conform lui Charlie*, in the same example, is a rather unnatural choice in the informal register to which the sentence seems to belong. Though acceptable in a more formal context, *conform* is replaced by *după* in more relaxed communication settings.

Grammatically, agreement is faulty both in terms of gender and in terms of number (in Romanian, nouns agree with their determiners and modifiers in gender, number and case): the Romanian word *mazăre* is a collective noun assimilated to the feminine gender, while here, it is accompanied by a masculine determiner – the numeral *doi*. The feminine form of the numeral would, however, not have saved the translation by itself, as *mazăre* requires a special partitive if it is to be turned into a countable noun: *un bob de* (en. *a grain of*). *Păstăi*, in the plural, is incorrectly suggested instead of the singular *păstaie*, which seems to be announced by the indefinite singular feminine article *o*. While this is of little relevance in the case of the idiomatic expression "like two peas in a pod", it may help when calibrating the translation of this phrase taken literally.

Ex 9 Well, according to Charlie, you and this Mrs Sparrow are like two peas in a pod.
Ei bine, conform lui Charlie, tu și această doamnă Sparrow sunteți ca doi mazăre într-o păstăi.

Some weakness of the machine translation platform is also revealed when dealing with polysemantic words (indicated in bold in Examples 10 to 16) or with longer and more complex sentences (though the asset of a neural translation system should have been exactly this – its ability to successfully deal with bigger language chunks). Google Translate is not always able to pick up the right, contextual meaning out of many of a word or phrase – it shows a tendency to rather suggest an equivalent for the most frequent, basic meaning of the source language lexical item, a choice which often results into the intended purpose of the original message being completely altered (in however, a grammatically correct sentence):

Ex 10 My mother walks out of the room when my father brings up **sports**.
Mama iese din cameră când tatăl meu face sport.
(My mother walks out of the room when my father does sport.)

Target text readers are left confused as a consequence of the same inappropriate choice of meaning from a range available to provide an equivalent for in examples 11 to 16 below:

Ex 11 would be off his **rocker**
ar fi în afara rockerului său (would be outside her rock player)

Ex 12 sleeping dogs **lie**
câinii adormiți să mintă (the dogs asleep should lie)

Ex 13 **nail** on the head
unghia de cap (nail of the head)

Ex 14 The woman **broke down** when the police told her that her son had died.
Femeia s-a defectat când poliția i-a spus că fiul ei a murit.
(The woman ceased to function when the police told her that her son had died.)

Ex 15 I accidentally **ran over** your bicycle in the driveway.
Ți-am alergat accidental peste bicicleta pe alee.
(approx. I accidentally went over your bicycle in the driveway while I was running.)

Ex 16 The police **set up** the car thief by using a hidden camera.
Poliția a pus la cale hoțul de mașini folosind o cameră ascunsă.
(The police masterminded the car thief using a hidden camera.)

Politically correct terms also represent a challenge for Google Translate, as it does not grasp their special euphemistic touch. Not being able to do that, it offers a word for word translation that results into awkward equivalents. In the particular case of Example 17, what was meant as a politically correct statement turns into the opposite in Romanian, as the feminine gender is implicitly assigned, via the adjective *grasă*, to the otherwise gender-neutral English personal pronoun "you":

Ex 17 Sorry, you're not fat, you're horizontally challenged.
Ne pare rău, nu ești grasă, ești provocată orizontal.
(Sorry, you are not fat, you are challenged in a horizontal manner.)

Random gender assignment for the same referent is also visible in examples like 18, where the gender-neutral "boss" is translated by a masculine Romanian noun – *șef*, but is then referred to by the feminine possessive pronoun *ei*:

Ex 18 Everyone seems to be trying to butter up the new boss hoping to become her favorite. *Toată lumea pare să încerce să-l ungă pe noul șef sperând să devină preferatul ei*.
(Everybody seems to be trying to butter up the new boss (masc.), hoping to become her favourite.)

We used Example 18 to illustrate the category of mixed errors as, besides the instance of random gender assignment that it contains, we also identified a lexical error in it – the metaphorical meaning of the slang phrase "to butter up" is unnaturally rendered by *să-(l) ungă* (en. lit. *to spread him*) which, if it were to signify approximately the same thing as the original, should have been *să-(l) ungă la suflet* (en. approx. *to spread something on his*

soul). An alternative translation solution could have been the metaphorical, slang as well, *să-(l) perie(ze)* (en. lit. *to brush him up*).

Finally, there are two instances in the everyday language sub-corpus which may represent spelling mistakes – *batonă* instead of *baton*, and *predez* instead of *predau*, but which may be also interpreted as morpho-syntactic errors. In the former case, a presupposed feminine form is used instead of the neuter (*batonă* does not actually exist in Romanian, but it has the form of most feminine nouns in this language). In the latter, the verb "a preda" – "to hand in" is conjugated following the wrong paradigm, most probably based on the model of other verbs of the 1st conjugation like "a lucra" (en. *to work*) – "lucrez" (en. *I work*), "a demola" (en. *to demolish*) – "demolez" (en. *I demolish*), "a picta" (en. *to paint*) – "pictez" (en. *I paint*). Below are the contexts in which these errors occur:

Ex 19 chocolate bar
*batonă** de ciocolată

Ex 20 I have to hand in my essay by Friday.
Trebuie să-mi predez eseul până vineri.*

Table 2 shows the errors in the translation from English into Romanian, by the Google digital tool, of the newspaper/ press releases language sub-corpora.

Table 2 Errors in the Google translation of newspaper/ press releases language samples (English into Romanian)

Error type	ST (English)	TT (Romanian)	Sub-corpus
lexico-semantic	Brent crude futures fluctuated Monday.	*La termen, brentul brut va fluctua luni.*	NP6
	It has long been regarded as a soft power superpower.	*A fost mult timp considerată o superputere de putere moale.*	NP3
	a landslide victory	*o victorie alunecătoare*	NP2
	some of the posts were more alarming	*unele dintre posturi au fost mai alarmante*	NP2
	The research released by reporting forum Stop AAPI Hate on Tuesday revealed nearly 3,800 incidents were reported.	*Cercetările lansate marți de forumul de raportare Stop AAPI Hate au dezvăluit marți că aproape 3.800 de incidente au fost raportate.*	NP1
	the ship (...) after running aground while entering the Suez Canal	*the ship (...) după ce s-a prăbușit în timp ce pătrundea în Canalul Suez*	NP6
	who accuse him of helping trash Brazil's international reputation	*care îl acuză că a ajutat la restabilirea reputației internaționale a Braziliei*	NP3
	One Chinese American woman	*O femeie americană chineză*	NP1
	Another woman, who's Filipino American	*O altă femeie, care este filipinez americană*	NP1
	a brawl about five blocks from the White House	*o bătaie la aproximativ cinci blocuri de Casa Albă*	NP2
	honking their horns	*claxonând din coarne*	NP2
	the ship, known as the Ever Given	*nava, cunoscută sub numele de Ever Date*	NP6
	to avoid the logjam at the canal	*să evite logjamul la canal*	NP6
	Brazil's shambolic response to coronavirus	*răspunsul shambolic al Braziliei la coronavirus*	NP3
	neopronoun; neopronouns, including ze/zir, xe/xim and fae/faer	*neopronumul; neopronome inclusiv ze/zir, xe/xim și fae/faer*	NP5

Error type	ST (English)	TT (Romanian)	Sub-corpus
morphological-syntactic	there are signs of similar alarm	*există semne de alarmă similar*	NP3
	making it by far the most deadly month of Brazil's 13-month epidemic	*făcându-l cea mai mortală lună a epidemiei de 13 luni a Braziliei*	NP3
	which 5 percent of the surveyed LGBTQ youths reported using exclusively	*care 5 la sută din tinerii LGBTQ chestionați au raportat că folosesc exclusive*	NP5
	The canal authority said maritime traffic will resume.	*Autoritatea canalului a declarat că traficul maritim va relua.*	NP6
	But to many of his most fervent supporters, these facts didn't matter, and still don't.	*Dar pentru mulți dintre cei mai fervenți susținători ai săi, aceste fapte nu au contat și încă nu au.*	NP2
	the result of land and ecological degradation	*rezultatul degradării terenurilor și ecologice*	NP4
	Close to 60,000 Brazilians are expected to die in March alone.	*Aproape 60.000 de brazilieni sunt așteptați să moară doar în martie.*	NP3
lexical-semantic + morphological-syntactic	anxieties over the global supply chain, which had already been impacted by the coronavirus pandemic	*anxietățile legate de lanțul global de aprovizionare, care fusese deja afectată de pandemia de coronavirus*	NP6
	Dozens of members (...) who (...) were among those who would later break into the US Capitol, joined the march.	*Zeci de membri (...) care (...) s-au numărat printre cei care ar intra ulterior în Capitolul SUA, s-au alăturat marșului.*	NP 1
	a man on the subway slapped my hands, threatened to throw his lighter at me	*un bărbat de la metrou mi-a dat o palmă, mi-a amenințat că aruncă bricheta asupra mea*	NP2
orthographic	Pollution has proved to be a pernicious challenge.	*Poluarea sa* [s-a] dovedit a fi o provocare periculoasă.*	NP4
	reinsurers' earnings	*câștigurile reasigurătorilor* [reasiguratorilor]*	NP6

A close analysis of the Google translation of the newspaper articles and releases by news agencies in our corpus reveals the fact that, like in the case of everyday language, most types of errors are lexico-semantic. These are of various kinds, as we are going to explain below.

Field-specific terminology is sometimes mistranslated. In Example 21, "Brent crude futures" is a business term meaning "Brent oil future contracts". Google Translate is not able to interpret it as such; it even considers the proper noun "Brent" a common one and offers a nonsensical translation: *La termen, brentul brut va fluctua luni.* (en. approx. *When due, the brent* will fluctuate on Monday*). In Example 22, the word-for-word translation of the specialized (metaphorical) term "soft power", used in politics to refer to the use of positive attraction and persuasion to achieve foreign policy objectives, also results into an awkward Romanian equivalent: *superputere de putere moale.*

Ex 21 Brent crude futures fluctuated Monday.
La termen, brentul brut va fluctua luni.

Ex 22 It has long been regarded as a soft power superpower.
A fost mult timp considerată o superputere de putere moale.

Similarly to what happens in the case of some polysemantic words in everyday speech, in this sub-corpora too, the Google digital translation platform selects the wrong meaning to consider for the Romanian equivalent. The

result is, once again, unnatural in the target language – either collocations or words that are discordant in the contexts in which they are employed. In Example 23, the meaning of "landslide" that is picked up by the machine is "collapse of a mass of earth" instead of "overwhelming majority of votes in an election". In Example 24, for "posts", the meaning based on which the Romanian counterpart is provided is "position in a company or organization", instead of "something that is published, announced or advertised". Meaning selection may have been appropriate for the verb "to release" in Example 25 – "make available to the public", but the Romanian verb "a lansa" (en. *to launch*), which may be used with the meaning of the English original in some contexts, is excluded from the collocation with the noun "cercetări" (en. *researches*). In this particular context, "cercetări" collocates with a phrase rather than with the single verb indicated by Google – "făcute publice" (en. *made public*). A redundancy error also occurs in Example 25 – the Romanian word *marți* for the English "Tuesday" is unnecessarily employed twice in the target sentence.

Ex 23 a landslide victory
o victorie alunecătoare

Ex 24 some of the posts were more alarming
unele dintre posturi au fost mai alarmante

Ex 25 The research released by reporting forum Stop AAPI Hate on Tuesday revealed nearly 3,800 incidents were reported.
Cercetările lansate marți de forumul de raportare Stop AAPI Hate au dezvăluit marți că aproape 3.800 de incidente au fost raportate.

Some of the English words are translated completely wrong, with very little possible justification for that beyond the algorithmic associations made by the neural translation system. This is the case of "running aground" translated as *s-a prăbușit* (en. *collapsed*) in Example 26 and "trash... Brazil's reputation" rendered into Romanian as *restabilirea... reputației Braziliei* (restoration ... of Brazil's reputation) in Example 27.

Ex 26 the ship (...) after running aground while entering the Suez Canal
the ship (...) după ce s-a prăbușit în timp ce pătrundea în Canalul Suez

Ex 27 who accuse him of helping trash Brazil's international reputation
care îl acuză că a ajutat la restabilirea reputației internaționale a Braziliei

There are cases when the same English word has a Romanian equivalent with multiple, context-dependent forms or several different, also context-dependent equivalents. Google Translate has demonstrated that it cannot always make the appropriate choice from among the options it may have in the case of such English words. When it cannot, the variants it suggests lack naturalness in Romanian. This is illustrated by *o femeie americană chineză* for "One Chinese American Woman" in Example 28, where the form of the Romanian adjectives indicating nationality should have been *americancă chinezoică*. Similarly, in Example 29, the correct Romanian form to be used istead of *filipinez americană* is *americancă filipineză*. Better still in both cases, a phrase should have been suggested: *americancă de origine chineză* (en. *American of Chinese origin*) and, respectively, *americancă de origine filipineză* (en. *American of Filipino origin*). In its turn, "anxieties", which may be translated as *anxietăți* in some contexts, should have more appropriately been rendered as *temeri* (en. *fears*) in Example 30. Also note in Example 30 the disagreement in gender of the neuter subject noun *lanțul* and the passive form of the predicate verb "a afecta" – *fusese... afectată*, which is gender-sensitive in Romanian.

Ex 28 One Chinese American woman
O femeie americană chineză

Ex 29 Another woman, who's Filipino American
O altă femeie, care este filipinez americană

Ex 30 anxieties over the global supply chain, which had already been impacted by the coronavirus pandemic
anxietățile legate de lanțul global de aprovizionare, care fusese deja afectată de pandemia de coronavirus

There are English words that do not mean the same thing in different geographical varieties of this language, a fact that sometimes puzzles Google Translate to the point that it offers the Romanian equivalent for the meaning that is not actually intended in the original. This is the case of the American English "blocks", in Example 31, wrongly taken and translated for its British English use by *blocuri* and the American English "horns", in Example 32, taken to belong to the British variety and faultily translated as *coarne*.

Ex 31 a brawl about five blocks from the White House
o bătaie la aproximativ cinci blocuri de Casa Albă

Ex 32 honking their horns
claxonând din coarne

Partial translation and borrowing are other causes of errors or difficult reception of the target text in English-into-Romanian translations by Google Translate. This is obvious in Example 33, where "Ever Given", the name of a ship, a compound proper noun, is half-translated, half-taken over with its original form – *Ever Date* (*Date* indicates that "Given" was interpreted by the machine as being a common noun for which the Romanian may be, indeed, *Date*; if this is true, the capital letter in the word remains puzzling). "Logjam" is a solid compound in English that refers to an agglomeration of logs on a river. "Shambolic" is a portmanteau word, a combination of "shambles" and "symbolic", meaning 'disorganized, messy, confused'. Both "logjam" and "shambolic", most probably unknown to the translation platform, have been taken over as such in the Romanian sentences (the former in an assimilated, articulated form), potentially hindering their clear reception.

Ex 33 the ship, known as the Ever Given
nava, cunoscută sub numele de Ever Date

Ex 34 logjam
logjamul

Ex 35 Brazil's shambolic response to coronavirus
răspunsul shambolic al Braziliei la coronavirus

Neologisms are also sometimes handled inefficiently by Google Translate – "neopronoun" and its plural, "neopronouns" are rendered into Romanian as the confusing *neopronumul* and *neopronome*, respectively.

Ex 36 neopronoun; neopronouns, including ze/zir, xe/xim and fae/faer
neopronumul; neopronome inclusiv ze/zir, xe/xim și fae/faer

In the morpo-syntactic category, there is also a variety of errors that have been identified.

Faulty agreement, mostly as far as gender is concerned, occurs in two instances: in Example 37, *similar*, for the English "similar", should have had the form *similare*, specific for the neuter plural, so as to agree with *semne* (en. *signs*) as its head-noun; in Example 38, the weak form of the 3rd person singular, masculine personal pronoun *-l* should have taken the form for the feminine, as it agrees with *lună* (en. *month*), the head-noun it anticipates.

Ex 37 there are signs of similar alarms
există semne de alarmă similar

Ex 38 making it by far the most deadly month of Brazil's 13-month epidemic
făcându-l cea mai mortală lună a epidemiei de 13 luni a Braziliei

Sometimes, the morphological features of either a source or a target language part of speech are misinterpreted or disregarded and the resulting translation solutions are, therefore, incorrect. Examples 39 and 40 illustrate this. In 39, the adverb "exclusively" is rendered as the adjective *exclusive*, while in 40, the modal verb "would", expressing willingness, determination, is interpreted as an auxiliary in the structure of the conditional. Consequently, "would … break" is translated as *ar intra*. In Example 40 as well, the proper noun "(US) Capitol" is taken to be a common noun and is translated by *Capitolul (SUA)* (Chapter (USA)). Oher times, the digital translation platform is unable to resort to transposition when, for example, the Romanian reflexive voice is needed for the English active – see Example 41, where "will resume" is translated by the active *va relua*, instead of *se va relua*. In English, if a verb present in a sentence needs to be repeated farther in that sentence, it may be replaced by its non-translatable corresponding auxiliary ("do", "be" or "have"), which is not possible in Romanian, where the full verb has to be visible in the surface structure every time this is needed. Google Translate disobeyed this rule and provided a translation for "don't" in Example 42 – the Romanian auxiliary *au* (have). Transitivity and the use of objects also proves to sometimes be problematic for Google Translate, as it happens in Example 43, where the Romanian equivalent for "threatened" – *a amenința*t, even if transitive, does not accept an indirect object in the dative. So, *mi-*, the weak form of the Romanian 1st person singular personal pronoun in the dative is incorrectly suggested here. Example 43 also contains two lexical translation errors: the preposition "on" in "on the metro" indicates that the person is in the underground vehicle, while its Romanian equivalent – *de la* rather suggests that the person works for the metro; "my hands" is omitted from the translation so that the sense of the original is (slightly) altered.

Ex 39 which 5 percent of the surveyed LGBTQ youths reported using exclusively
care 5 la sută din tinerii LGBTQ chestionați au raportat că folosesc exclusive

Ex 40 Dozens of members (…) who (…) were among those who would later break into the US Capitol, joined the march.
Zeci de membri (…) care (…) s-au numărat printre cei care ar intra ulterior în Capitolul SUA, s-au alăturat marșului.

Ex 41 The canal authority said maritime traffic will resume.
Autoritatea canalului a declarat că traficul maritim va relua.

Ex 42 But to many of his most fervent supporters, these facts didn't matter, and still don't.
Dar pentru mulți dintre cei mai fervenți susținători ai săi, aceste fapte nu au contat și încă nu au.

Ex 43 a man on the subway slapped my hands, threatened to throw his lighter at me
un bărbat de la metrou mi-a dat o palmă, mi-a amenințat că aruncă bricheta asupra mea

At a higher level of complexity, there are errors which are generated by misinterpretation of the structure of certain phrases. In Example 44, the noun "degradation" is accompanied by two coordinated modifiers – "land and ecological". However, only one of them – "land" is interpreted as what it actually is, the other one being perceived as an adjective in isolation from the phrase. The Nominative + infinitive construction in Example 45, in which the two elements are connected by a verb in the passive voice – "Brazilians are expected to die" is not correctly understood as containing a noun in a subject-predicate relationship with the infinitive. Instead, the subject is believed to be the subject of the passive voice predicate, which triggers a translation whose meaning is detached from the one intended.

Ex 44 the result of land and ecological degradation
rezultatul degradării terenurilor și ecologice

Ex 45 Close to 60,000 Brazilians are expected to die in March alone.
Aproape 60.000 de brazilieni sunt așteptați să moară doar în martie.

Orthographic errors are scarce in the translated newspaper articles sampled in our corpus. They are illustrated in Examples 46 and 47, where incorrect spelling is marked with (*) and the correct one is provided within square brakets:

Ex 46 Pollution has proved to be a pernicious challenge.
Poluarea sa [s-a] dovedit a fi o provocare periculoasă.*

Ex 47 reinsurers' earnings
câștigurile reasigurătorilor [reasiguratorilor]*

Conclusions

For developers and users of the Google Translate service, it is most obvious that it has progressed during the seventeen years since it was launched. However, it is equally obvious that it still has limitations that prevent it from providing human-like, first-rate translations for all languages and text genres that it is made to work with.

Our small-scale study demonstrated that, as far as the English-Romanian language pair is concerned, in the particular case of everyday language, mostly used in informative, short newspaper articles and news releases as well, the outcome of the machine-driven translation process still calls for human translators' intervention to bring it up to a high-quality standard. However, in the case of texts belonging to the two genres selected for this research, the Google translations we obtained lack functionality only in very few cases that cover rather limited stretches of language. If we agree with Zetzsche's (2010) perspective that the quality of a translation should be assessed based on what is actually expected of that particular translation, in other words, that, "since translation quality is very abstract and arguable [...] the only relevant measure for translation is usefulness", we can conclude that Google Translate can generally produce useful, useable, functional, reasonably fluent and accurate translations of everyday English into Romanian, .

The errors that are still visible in the target Romanian texts do not generally hinder understanding of these texts. They should, nevertheless, be called to the attention of both human translators (who may find them useful especially for the post-editing of machine translation) and platform designers (who may refer to them to improve the performance of the Google neural translation system). To round our research off, the following types of English-into-Romanian translation errors are brought to the fore for the everyday and newspaper/ news releases language, as sub-categories of Keshavarz's (1999) general error classes:

- In the lexico-semantic category:
  - incorrect choice of equivalents both for monosemantic and for polysemantic words;
  - incorrect choice of equivalents for words with a metaphorical/ euphemistic meaning;
  - incorrect choice of equivalents for domain-specific terms;
  - incorrect (direct) translation of idiomatic phrases;
  - incorrect interpretation of geographical varieties of English;
  - incorrect interpretation of neologisms;
  - incorrect choice of register.
- In the morphological-syntactic category:
  - incorrect agreement in number, gender and case;
  - random gender assignment;
  - incorrect inflection and conjugation;
  - misinterpretation of features of certain morphological categories;
  - misinterpretation of phrase structure.
- Spelling errors were rarely encountered in the Romanian texts. They mainly consist of:

- misspelling of words;
- incorrect use of the hyphen.

We look at the present research as a starting point in carrying out similar analyses on other genres so as to draw some conclusions on genre-specific and generalized errors that Google Translate still makes when an English input has to be turned into Romanian. A comparison between translations of the same texts provided at different moments in time would also be useful to draw as it would serve as a barometer of the system functionality improvement or stagnation, as the case may be. Further investigation that would allow us to rank the errors as far as their gravity is concerned would also be welcome.

Acknowledgements

We thank Ciprian Meteș, research assistant at the CODHUS research centre, West Univesity of Timișoara, Romania, for data collection and his insightful observations.

Conflict of Interest

The authors declare no conflict of interest regarding the publication of this article.

References

1 Ageeva, E., Tyers, F., Forcada, M., & Perez-Ortiz, J. (2015). Evaluating machine translation for assimilation via a gap-filling task. Proceedings of the Conference of the European Association for Machine Translation, Antalya, Turkey, 137-144. Retrieved September 8, 2023, from https://aclanthology.org/W15-4918.pdf

2 Bapna, A., Caswell, I., et al. (2022). Building machine translation systems for the next thousand languages. Cornell University, arXiv preprint. arXiv:2205.03983, 1-77. https://doi.org/10.48550/arXiv.2205.03983

3 Brezina, V., Weill-Tessier, P., & McEnery, A. (2021). #LancsBox v. 6.x. [software package] http://corpora.lancs.ac.uk/lancsbox/index.php

4 Chatzikoumi, E. (2020). How to evaluate machine translation: a review of automated and human metrics. Natural Language Engineering, 26(2), 137-161. https://doi.org/10.1017/S1351324919000469

5 De la Cruz-Cabanillas, I. D. L., & Tejedo-Martínez, C. (2016). The error analysis approach for the assessment of automatic translation. Lingwistyka Stosowana, 16(1), 1-9. https://doi.org/10.32612/uw.20804814.2016.1.pp.1-9

6 Dumitran, A. (2021). Translation error in Google Translate from English into Romanian in texts related to coronavirus. International Scientific Conference eLearning and Software for Education (vol. 2, pp. 37-43). https://doi.org/10.12753/2066-026X-21-078

7 Keshavarz, M. H. (1999). Contrastive Analysis and Error Analysis. Tehran: Rahnama Press.

8 Koponen, M. (2010). Assessing machine translation quality with error analysis. MikaEL: Electronic Proceedings of the KäTu Symposium on Translation and Interpreting Studies, 4. Suomen kääntäjien ja tulkkien liitto. Retrieved June 15, 2023, from https://sktl-fi.directo.fi/@Bin/40701/Koponen_MikaEL2010.pdf

9 McEnery, T., & Hardie, A. (2012). Corpus Linguistics: Method, Theory and Practice. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511981395

10 Popović, M. (2018). Error classification and analysis for machine translation quality assessment. In J. Moorkens, S, Catsilho, F. Gaspari & S. Doherty (Eds.). Translation Quality Assessment. From Principles to Practice. Machine Translation: Technology and Applications Series (vol 1, pp. 129-158). Cham: Springer. https://doi.org/10.1007/978-3-319-91241-7_7

11 Sanders, G., Przybocki, M., Madnani, N., & Snover, M. (2011). Human subjective judgments. In J. Olive, J. McCary & C. Christianson (Eds.). Handbook of Natural Language Processing and Machine Translation. DARPA Global Autonomous Language Exploitation (pp. 806-807). New York: Springer.

12 Secară, A. (2005). Translation evaluation – a state-of-the-art survey. Proceedings of the eCoLoRe/MeLLANGE Workshop, Leeds, 39-44.

13 Snover, M., Madnani, N., Dorr, J. B., & Schwartz, R. (2009). Fluency, adequacy or HTER? Exploring different human judgements with a tunable MT metric. Proceedings of the Fourth Workshop on Statistical Machine Translation, StatMT '09, 259-268. Stroudsburg: Association for Computational Linguistics. Retrieved May 26, 2023, from https://doi.org/10.3115/1626431.1626480

14 Zetzsche, J. (2010). Pondering and wondering. Translation Journal, 14(1), January 2010. Retrieved August 31, 2023, from https://translationjournal.net/journal/51pondering.htm

15 Way, A. (2018). Quality expectations of machine translation. In J. Moorkens, S. Castilho, F. Gaspari & S. Doherty (Eds.). Translation Quality Assessment. From Principles to Practice. Machine Translation: Technology and Applications Series (vol.1, 159-178). Cham: Springer. https://doi.org/10.1007/978-3-319-91241-7_8

16 Wise, J. (2023). Google Translate users: how many people use it in 2023? Earthweb. Retrieved September 8, 2023, from https://earthweb.com/how-many-people-use-google-translate/

Sources

1 NP1: Yam, K. (2021). There were 3,800 anti-Asian racist incidents, mostly against women, in past year. NBC News. Retrieved March, 2023, from https://www.nbcnews.com/news/asian-america/there-were-3-800-anti-asian-racist-incidents-mostly-against-n1261257

2 NP2: Sardarizadeh, S., Lussenhop, J. (2021). The 65 days that led to chaos at the Capitol. BBC News. Retrieved March, 2023 from https://www.bbc.com/news/world-us-canada-55592332

3 NP3: Philipps, T., Goñi, U., Parkin Daniels, J. (2021). The heart of darkness: neighbours shun Brazil over Covid response. The Guardian. Retrieved March, 2023, from https://www.theguardian.com/global-development/2021/mar/30/neighbors-shun-brazil-covid-response-bolsonaro

4 NP4: Lee Myers, S. (2021). The worst dust-storm in a decade shrouds Beijing and Northern China. The New York Times. Retrieved March, 2023, from https://www.nytimes.com/2021/03/15/world/asia/china-sandstorm.html

5 NP5: Venkatraman, S. (2020). Beyond 'he'and 'she': 1 in 4 LGBTQ youths use nonbinary pronouns, survey finds. NBC News. Retrieved March, 2023, from https://www.nbcnews.com/feature/nbc-out/beyond-he-she-1-4-lgbtq-youths-use-nonbinary-pronouns-n1235204

6 NP6: Lee, N. Y. (2021). Cargo ship blocking the Suez is partially floated, says Suez Canal Authority. CNBC. Retrieved March, 2023, from https://www.cnbc.com/2021/03/29/cargo-ship-blocking-suez-canal-has-been-refloated-says-inchcape.html

Santrauka

Loredana Pungă, Ionela Manda, Mădălina Chitez

Kaip gerai *Google Translate* moka rumunų kalbą? Tekstynu grįsta skirtingų žanrų tekstų vertimo iš anglų kalbos į rumunų kalbą klaidų analizė

Siekiant nors ir nedideliu mastu kompensuoti anglų-rumunų kalbų įvairių žanrų tekstų mašininio vertimo tyrimų trūkumą, šioje atvejo studijoje nagrinėjamas dviejų skirtingų, bet glaudžiai susijusių žanrų tekstų – buitinio teksto ir laikraščių bei naujienų pranešimų kalbos – vertimas iš anglų kalbos į rumunų kalbą pasitelkiant „Google" vertėją. Siekiama pateikti vertimo klaidų rumunų kalbos tekstuose ir vertimo kokybės analizę. Šiam tikslui pasiekti atliekama vertimo klaidų analizė, remiantis Keshavarz'o (1999) klaidų analizės modeliu ir sukuriami abiejų nagrinėjamų žanrų kalbinių klaidų profiliai. Nustatytos klaidos aptariamos ir iliustruojamos nedidelės apimties tekstyno pavyzdžiais. Kadangi vertimo klaidos, darančios įtaką vertimo kokybei, tiesiogiai priklauso nuo „Google" vertimo įrankio galimybių, šio straipsnio išvados gali būti aktualios minėto įrankio kūrėjams. Jie gali susidaryti aiškesnį vaizdą apie jo stipriąsias ir silpnąsias puses ir pasiūlyti būdų, kaip jį patobulinti, kad galiausiai galima būtų užtikrinti aukštesnę vertimo kokybę anglų–rumunų kalbų poroje. Šis tyrimas taip pat gali padėti atkreipti vertėjų dėmesį į sritis, kurios tokiomis aplinkybėmis galėtų kelti problemų atliekant postredagavimą.

About the Authors

LOREDANA PUNGĂ

Professor, West University of Timișoara, Romania

Research Interests
Translation studies, applied linguistics, cognitive linguistics

Address
Blvd. Vasile Parvan 4, Timișoara 300223, Romania

E-mail loredana.punga@e-uvt.ro

Orcid iD 0000-0003-0544-237X

IONELA MANDA

PhD student, CODHUS, West University of Timișoara, Romania

Research Interests
Corpus linguistics, applied linguistics

Address
Blvd. Vasile Parvan 4, Timișoara 300223, Romania

E-mail ionela.manda@e-uvt.ro

MĂDĂLINA CHITEZ

Senior researcher, CODHUS, West University of Timișoara, Romania

Research Interests
Applied corpus linguistics, digital humanities, academic writing

Address
Blvd. Vasile Parvan 4, Timișoara 300223, Romania

E-mail madalina.chitez@e-uvt.ro

Orcid iD 0000-0001-9005-3429