Current Shortcomings of Machine Translation in Spanish and Bulgarian Vis-à-vis English

Travis Sorenson
University of Central Arkansas
tsorenson@uca.edu

Abstract

In late 2016, Google Translate (GT), widely considered a machine translation leader, replaced its statistical machine translation (SMT) functions with a neural machine translation (NMT) model for many large languages, including Spanish, with other languages following thereafter. Whereas the capabilities of GT had previously advanced incrementally, this switch to NMT resulted in seemingly exponential improvement. However, half a dozen years later, while recognizing GT's usefulness, it is also imperative to systematically evaluate ongoing shortcomings, including determining which challenges may reasonably be presumed as superable over time and those which, following a multiyear tracking study, prove unlikely ever to be fully resolved. While the research in question principally explores Spanish-English-Spanish machine translation, this paper examines similar problems with Bulgarian-English-Bulgarian GT renditions. Better understanding both the strengths and weaknesses of current machine translation applications is fundamental to knowing when such non-human natural language processing (NLP) technology is capable of performing all or most of a given task, and when heavy, perhaps even exclusive human intervention is still required.

Keywords: Bulgarian, English, Google Translate, machine translation, Spanish

1 Theoretical introduction and historical overview

The genesis of this study lies in events that, while years in the making, came to light fully in late 2016, when programmers behind the scenes switched the online machine translation service Google Translate (GT) from one employing statistical machine translation (SMT) to one relying on the company's newly completed neural machine translation (NMT) system (Lewis-Kraus, 2016). Rather than featuring different modules, NMT utilizes a single, streamlined system that contains only an encoder, which analyzes the training data (mostly bilingual corpora), and a decoder, which applies this analysis to a new source-language text and renders it into the chosen target language. While the encoder assigns individual words and other features numerical qualities, the decoder considers texts to be translated at the full sentence level, rather than according to separate words or phrases as with the SMT models (Poibeau, 2017: 185). This seeming simplicity should not obscure the fact that NMT is not only extremely complex, but, given that the representation of the data is strictly numerical, it is not completely understood even by those who have written the algorithms leading to the vectors of numbers involved in the work of encoding the bilingual texts (193). While NMT can compete with human translators on tasks involving highly repetitive structures (e.g. legal documents, economic texts), less common and more creative, more novel content can lead to serious meaning errors. In other words, adequacy may suffer even if the fluency of the resulting translation may be acceptable. The main cause of this difficulty on the part of computers to engage successfully in natural language processing (NLP) is ambiguity (lexico-semantic, morphosyntactic, etc.) (Koehn, 2020: 37). A number of examples displaying this phenomenon are considered in this paper.

2 Overview of research

The purpose of the larger research project – based mainly on English and Spanish – is to determine not only what NMT can and cannot do, both generally and specifically, but also what improvements might occur over the next five years or so. As this research focuses largely on the written word, a thorough assessment of these matters requires a systematic evaluation of the different categories involved, namely expository writing, descriptive/narrative writing, and persuasive writing via texts from numerous subcategories in each case. This allows for methodical contrasting and comparing of the results yielded by GT, some of which are unique to the relationship between English and Spanish, whereas others have features that can be extrapolated to other languages, including Bulgarian.

3 Presentation, discussion, and analysis of research data

Initial work drawing on material from the above-mentioned categories will now allow for a discussion of GT results stemming from a variety of texts. While GT is capable of performing felicitous translations in many content areas, and while it has arguably improved vastly in all areas since the switch to NMT beginning in 2016, many renditions continue to be problematic to one degree or another. While many such instances are given throughout this paper, it is appropriate at this point to share two such examples (i.e. one that is considered to be a suitable, even excellent translation by GT, and another that is decidedly flawed). Both happen to be from English to Spanish. The first is from an economic report published by the business and finance website *cnbc.com* (Fitzgerald and Stevens, 2021):

(1) July’s Consumer Price Index released Wednesday showed prices jumped 5.4% since last year, compared to expectations of 5.3%, according to economists surveyed by Dow Jones. The government said CPI increased 0.5% in July on month-to-month basis.

GT (4 Oct 2021):

El índice de precios al consumidor de julio publicado el miércoles mostró que los precios subieron un 5,4% desde el año pasado, en comparación con las expectativas del 5,3%, según economistas encuestados por Dow Jones. El gobierno dijo que el IPC aumentó un 0,5% en julio mes a mes.

The GT rendition of passage (1) into Spanish represents an arguably flawless translation, including several important but somewhat subtle details. For instance, the initials CPI for Consumer Price Index have been converted, appropriately, to IPC (***í****ndice de* ***p****recios al* ***c****onsumidor*). Next, whereas in English no article is used before the expression of percentages, GT inserted an indefinite article in one case (*un 5,4%*), and a definite one in the other (*el 5,3%*). The use of either or both is more common than not in authentic Spanish. Finally, while the decimal separator employed in English is the period, most South American countries use the comma, which is also the case in Spain, which is part of the European Union. Apropos of that, the main reason for the accuracy of this translation, including all the intricacies mentioned, surely lies in the fact that myriad such texts that have been translated between English and Spanish and vice versa – such as those related to the European Parliament’s Committee on Economic and Monetary Affairs (ECON), which are found in the Europarl parallel corpus, to which GT has access.

The second example is a set of simple questions that one may easily presume would pose no great difficulty for GT:

(2) How are you, Dad?

How are you, Father?

How are you, Mom?

How **are you**, Mother?

GT (26 Mar 2022):

¿Cómo estás, papá?

¿Cómo estás, padre?

¿Cómo estás mamá?

¿Como ***está tu*** *madre?*

Whereas the first two translated sentences in example (2) are in no way problematic, the final two have a minor and then major errors. While *¿Cómo estás mamá?* inexplicably lacks the comma present in the two previous sentences, it is a detail that does not seriously impede understanding, essentially continuing to pose the same question. The final sentence, in contrast, suffers a catastrophic semantic change with a shift from second to third person: *¿Como está tu madre?* means ‘How **is your** mother?’, not ‘How **are you**, Mother?’

The following sections explore various issues, organized by common themes, that arise in GT renditions of original texts in different combinations of Spanish, English, and Bulgarian.

3.1 Pronoun-dropping and pronoun confusion between animate and inanimate objects

Whereas pronoun-dropping rarely occurs in English, it is quite common in many other languages, including Spanish and Bulgarian. In these pro-drop languages, other context markers, such as verb conjugations, serve to supplant much of the information carried in the missing pronouns, especially subject pronouns. However, since the context unavoidably becomes more implicit in the absence of the explicit pronouns, ambiguity unavoidably results. Although this type of situation is routinely processed without difficulty by humans, translation platforms such as GT are prone to significant meaning errors under the same conditions, as the following cases from Spanish and Bulgarian into English aptly demonstrate.

Writing about her experience covering the election of Pope Francis, Argentine journalist Elisabetta Piqué wrote the following in her book *Francisco: vida y revolución* (2014):

(3) *Lo recuerdo bien. **Estaba** en la plaza, embarazada de mi primer hijo de 6 meses, Juan Pablo.*

GT (8 Apr 2022):

‘I remember it well. **She** was in the plaza, pregnant with my 6-month-old first child, Juan Pablo.’

Author’s translation:

‘I remember it well. I was in the plaza, six months pregnant with my first child, Juan Pablo.’

Beyond committing the also serious error of stating that a fetus at the sixth-month stage of pregnancy is in reality a sixth-month-old child (a miscalculation of approximately nine months), GT takes what is clearly (to a human reader) a first-person reference and turns it into a third-person one. In the imperfect aspect of the Spanish past tense, the conjugation *estaba* (<infinitive *estar* ‘to be’) corresponds to various potential subjects: ‘I’ (*yo*), ‘he’ (*él*), ‘she’ (*ella*), ‘you’ (formal: *usted*), and ‘it’ (Ø). However, since the initial sentence was ‘I remember it well,’ it is clear that the one following also continues with the first person: ‘**I** was in the plaza…’ Not only would the GT rendition indicate that the father of the expected child was the one narrating and referring to the mother in the third person, but if such were the case the writer would almost surely have used an overt pronoun to make this abundantly clear: ***Ella estaba*** *en la plaza…* GT, processing largely at the sentence level, has no intersentential context on which to rely. Curiously, if only the segment *Estaba en la plaza* is processed, GT yields ‘**I** was in the square.’ It is likely, therefore, that the feminine word *embarazada* ‘pregnant’ incorrectly triggered a feminine pronoun: ‘she.’

A similar phenomenon can be witnessed with pronoun-dropping in Bulgarian, such as in the following pair of similar examples, the counterparts of which are also considered in Spanish:

(4) Вижда**м** **я**. Идв**а**.[1]
(I) see **her**. **(She)** comes.

(5) Вижда**м** **го**. Идв**а**.
(I) see **him**. **(He)** comes.

GT (5 Mar 2022):

‘I see **her**. **It’s** coming.’

‘I see **it**. **It’s** coming.’

(6) ***La*** *ve**o**.* *Vien**e**.*
Her **(I)** see. **(She)** comes.

(7) ***Lo*** *ve**o**.* *Vien**e**.*
Him **(I)** see. **(He)** comes.

[1] As the author of this paper does not speak Bulgarian, Google Translate was used to aid in the creation of some of the source-language sentences analyzed herein.

GT (8 Apr 2022):

'I see **<u>her</u>**. **<u>She</u>** comes.'

'I see. Comes.'

In the initial two-word sentence in both pairs of examples, it must be supposed that the speaker is referencing the sighting of a man and then a woman, respectively, as the literal gloss indicates. It should also be understood, in the case of the Bulgarian examples, that this language does not have infinitive verb forms; the first person singular conjugation in the present is therefore employed to refer to a verb, after which, as occurs in Spanish, the endings change for other persons and according to tense and aspect (see Table 1 for Bulgarian 'see'). Nevertheless, (4) and (5) both begin with a verb whose ending, in context, clearly refers, if only by default, to the first person singular. Use of the overt pronoun аз 'I' is unnecessary. However, in the second sentence of each example (a one-word verb phrase), the meaning is only implicit, as the verb form идва can mean 'he/she/it is coming.' The intent is clear to the speaker, but not to GT, which in both renditions has opted for the impersonal 'it.' Regarding example (6), even though the focus of GT's analysis tends to skew heavily to individual sentences, it seems that the presence of feminine *la* in the first sentence aided its correct choice of 'She' in its rendition of the second. However, not knowing if *lo* referred to 'him' or 'it,' GT omitted both in its version of (7), leaving a first sentence that lacks the needed object pronoun and a second one that is incomplete.

English	**Bulgarian**
'I see'	аз вижда**м**
'you see' (sing.)	ти вижда**ш**
'he sees'	той вижд**а**
'she sees'	тя вижд**а**
'it sees'	то вижд**а**
'we see'	ние вижда**ме**
'you see' (plur.)	вие вижда**те**
'they see'	те вижда**т**

Table 1: Present tense of Bulgarian verb 'see' (виждам) with overt subject pronouns

As manifested in example (5) in Bulgarian and (7) in Spanish, it is not only the omitted subject pronouns that can prove problematic for machine translation, but also the ambiguity of the clitics that are *not* left out. Whereas English features the unambiguous direct object pronouns 'me', 'you', 'him', 'her', 'it', 'us', and 'them', Spanish and Bulgarian both have counterparts of these pronouns that are clear in some instances and ambiguous in others. In Spanish, the equivalent of 'me', 'you' (sing. informal), 'us', and 'you' (plur. informal) are *me, te, nos, os,* all of which are distinct and therefore straightforward. However, 'him', 'her', 'you' (sing. formal), and 'it' can all be expressed *lo* or *la*, depending on gender, while 'them' and 'you' (plur. formal) are similarly either *los* or *las*. Faced with this uncertain situation in example (7), GT, as noted above, offered no equivalent pronoun at all in English, leaving only 'I see,' an intransitive verb use despite the fact that the sentence calls for a transitive construction, whether it be 'I see him' or 'I see it.' In Bulgarian, there is also some overlap in direct object pronouns, though it is limited to third person singular forms: 'him' and 'it' (masc./neut.) are both го, and 'her' as well as 'it' (fem.) are я (see Table 2 for all forms). In example (5), GT incorrectly selected inanimate 'it' in lieu of animate 'him,' even though in (4) it correctly chose animate 'her.' Nevertheless, if GT tends to render я as 'she' in all or most instances involving this type of ambiguity, eventually it will err – as it does in the next example – when this pronoun refers to an inanimate object that is assigned the female gender, the case with many Bulgarian nouns that end in –а and –я, such as маса 'table' in the following example given by Leafgren (2011: 74):

(8) Това е новата ни **<u>маса</u>**.
This is new our **<u>table</u>**.
Татко иска да **<u>я</u>** поставим
Father wants (aux.) **<u>it</u>** put
в ъгъла в
in (the) corner in (the)
кухнята.
kitchen.

'This is our new table. Dad wants us to put **<u>it</u>** in the corner in the kitchen.'

GT (7 Apr 2022):

'This is our new table. Dad wants us to put **<u>her</u>** in the corner of the kitchen.'

English	**Bulgarian (long)**	**Bulgarian (short)**
'me'	мен	ме
'you' (sing.)	Теб	те

‘him’	Него	го
‘her’	Нея	я
‘it’ (mas./neut.)	Него	го
‘it’ (fem.)	Нея	я
‘us’	Нас	ни
‘you’ (plur.)	Вас	ви
‘them’	Тях	ги

Table 2: Bulgarian accusative case (direct object) pronouns

3.2 Difficulties related to the gender of nouns and adjectives

As seen in the previous section, the use of certain third-person accusative pronouns in both Spanish and Bulgarian depends on the gender of either the person or the inanimate object that they modify. In both languages, gendered nouns themselves (and modifying adjectives) can also lead to difficulties for GT when ambiguities related to them arise in complex source-language material. In this regard, Koehn (2020: 7) proposes the following sentence in English, which is then translated into Spanish and Bulgarian, respectively:

(9) ‘Whenever I visit my uncle and his daughters, I can't decide who is my **favorite cousin**.’

GT (8 Apr 2022):

Cada vez que visito a mi tío y a sus hijas, no puedo decidir quién es mi ***primo favorito.***

Винаги, когато посещавам чичо си и дъщерите му, не мога да реша кой е **любимият** ми **братовчед**.

While a human has little problem with the logical deduction that the daughters of the uncle are by necessity female cousins, the link is not explicit enough for GT to avoid falling into the trap, which is set by the fact that English has no endings or any other morphological markings that render nouns and adjectives inherently masculine or feminine. As a result, in each instance, both the noun and its accompanying adjective were rendered in masculine form in the translation. In Spanish, ‘female cousin’ is ***prima*** rather than ***primo***, and the single feminine form of ‘favorite’ is ***favorita***, not ***favorito***. The same order of correct results in Bulgarian is **братовчедка** rather than **братовчед**, and **любимата** instead of **любимият**.

3.3 Lexical differences by regional dialect and the effects of homonymia

An important part of translation entails being able to insert the target language into its appropriate place in terms of culture and geography, a subfield of the discipline called localization. An essential element of this effort has to do with the suitable choice of specific vocabulary. If, for instance, a text in German about a *wohnung* were to be rendered into English, the translator would need to consider not only the target language but the pertinent dialect thereof. For a British audience the term ‘flat’ would be most appropriate, while US readers would identify with ‘apartment.’ In Spanish, at least three terms suggest themselves depending on the country or region: *piso* in Spain, *departamento* in Argentina, and *apartamento* in most of the rest of the Spanish-speaking world. The following examples, one from English to Spanish and the other in reverse order, are from the culinary world and show the importance of having certain lexical expertise in Spanish, a language particularly rich in synonym usage:

(10) ‘It is a common practice to sauté **mushrooms** in **butter**.’

GT (8 Apr 2022):

Es una práctica común saltear los **champiñones** en **mantequilla**.

(11) *Es más fácil tomar la soda con un* ***carrizo****.*

GT (9 Apr 2022):

‘It's easier to drink soda with a **reed**.’

The GT rendition of example sentence (10) would serve well in many Spanish-speaking countries, including certain large ones with high populations such as Mexico and Spain. In others, however, at least one of the food words would be uncommon to point of near non-existence. In Central American countries such as Honduras, Costa Rica, and Panama, the dominant term for ‘mushrooms’ is *hongos*. In Argentina, Uruguay, and Paraguay, the nearly universal term for ‘butter’ is *manteca*, despite that fact that in most other countries this name refers to ‘lard.’ Sentence (11) is one that could be heard throughout Panama, the only country were the default term for ‘drinking straw’ is *carrizo*, which, while it does mean ‘reed’ in other dialects, is employed metaphorically in this Central American country to denote a manmade hollow tube for sipping liquids. If Panamanians need to refer to a 'reed,' they have available for this purpose the word *caña* (which, in turn, is at times used in Perú not for a stem from the plant kingdom but, again, for 'drinking straw,' though the diminutive *cañita* is much more common for this purpose).

While national boundaries can often determine word usage, such as *carrizo* for 'drinking straw' in Panama alone, in some larger countries there may well be various intranational regions for a number of lexical items. For instance, in Spain, while speakers in nearly all areas of the country employ the term *judía* or *judía verde* to denote the 'green been,' in parts of the north, including the Basque Country, the term *vaina* prevails. Something similar is seen in Bulgaria, this time in the animal rather than the plant kingdom, and between the eastern and western zones of the country. Whereas the lexemes *gato* and 'cat' are universal in all dialects of Spanish and English, respectively, for the domesticated feline, Garavalova (2020) – referencing the *Bulgarian Etymological Dictionary*/Български етимологичен речник (BER, 1986) and the *Bulgarian Dialect Atlas*/Български диалектен атлас (BDA, 2001) – asserts that while speakers in the eastern two-thirds of Bulgaria tend to utilize the name **котка** (<*kotja 'female cat' <кот <Proto-Slavic *katъ <Latin *cattus*), in the western third or so of the country it is not uncommon to hear the term **мàчка** (etymology uncertain, but shared with Serbian in Cyrillic form and as *mačka* in Croatian, Slovak, and Slovenian with the same pronunciation) (104-106). Of interest, then, is the GT rendition of the following sentence:

(12) 'I don't like this **cat**.'

GT (9 Apr 2022):

Не харесвам тази **котка**.

If the target audience were speakers in central and eastern Bulgaria, the selection of **котка** would be optimal. If however, the intended group were those in the west, including the capital of Sofia, an acceptable localized rendition for many would be: Не харесвам тази **мàчка**. It is presumed, nevertheless, that most if not all speakers in western Bulgarian would understand both names – particularly if **котка** is the more normative of these two lexemes – though this is not always the case with dialectally determined vocabulary, especially with larger languages spread over wide expanses of the globe and several countries, such as English and Spanish.

A return to food vocabulary in the following Spanish-to-English translation will help to illuminate another issue that can arise with the GT renditions of texts involving dialectally based terminology.

(13) ***Las manías*** *son caras ahora mismo. Las almendras cuestan menos.*

GT (8 Apr 2022):

'**Crazes** are expensive right now. Almonds cost less.'

Author's translation:

'Peanuts are expensive right now. Almonds cost less.'

One immediately notices an incongruence in the machine translation, as 'crazes' do not generally carry a price tag and have precious little to do with 'almonds.' If, however, it is realized that both sentences in the original text concern a type of 'nut,' a bit of research on the matter can lead to a lexical solution in English. While the Náuhatl-derived *cacahuate* is the principle term for 'peanut' in Mexico, in the Caribbean, much of Central America, and all of South America, the dominant name is *maní* (plural *manís* or *maníes*), from the now-extinct Amerindian language Taíno. Yet only in Guatemala does one hear the altered form *manías*, as used in sentence (13), which is why GT failed to recognize the term's true meaning and translated it as 'crazes,' since in other contexts Spanish *manía* can signify 'mania,' a word denoting 'madness' in English that entered both languages via Latin from the earlier Greek. This means that the two cases of *manía(s)* in Spanish are homonyms: lexemes with the same spelling, the same pronunciation, but different meanings (typically with two different etymologies).

The difficulties presented at times by homonymia are not unique to Spanish; the following pair of examples shows that the phenomenon can also occur in English and Bulgarian, though it appears to be much more common in the former than the latter:

(14) 'The dog was old and sick; its **bark** was very weak.'

GT (23 Dec 2021):

El perro era viejo y estaba enfermo; su ***corteza*** *era muy débil.*

(15) ‘I like the feel of this **scythe**.’

GT (8 Apr 2022):

Харесва ми усещането за тази **коса**.

In the original text of sentence (14), the English word ‘bark’ obviously refers to the sound emanating from the dog’s mouth. The Spanish equivalent of this, however, is *ladrido*. The lexeme given by GT, *corteza*, refers to a protective outer layer of vegetable matter that constitutes the ‘bark’ of a tree. Moving to Bulgarian, the GT rendition from English of sentence (15) is only problematic if the reader does not understand the context, which could be that an individual has mentioned the need to find a useful tool for cutting grass or harvesting grain by hand. In isolation, however, **коса** could be understood as the homonym ‘hair.’ A person who does not often work with farm implements might well understand the word in its agricultural context and yet go years without encountering it in this setting. In contrast, one may refer to ‘hair’ on a weekly if not daily basis. This is surely the reason for the fact that GT, when processing a back translation of the Bulgarian sentence into English, opts for ‘hair’ as the equivalent of the term in question.

(16) Харесва ми усещането за тази **коса**.

GT (8 Apr 2022):

‘I like the feel of this **hair**.’

3.4 Additional examples in English and Spanish

Whereas many of the examples of problematic GT renditions shown to this point in the paper have included issues that one might find in relation to Bulgarian, there are myriad others that may pertain less or not at all to this language. If linguists of any native language are to grasp more fully the challenges still presented by machine translation, it is ultimately necessary that they expose themselves to such phenomena in multiple tongues, not to mention the various dialects of each. For instance, examples (15) and (16) above concerning the use of коса to denote both ‘hair’ and ‘scythe’ appears to be a rather rare instance of homonymia in Bulgarian. In contrast, the use of identically spelled and pronounced words in both Spanish and English is quite common. As a mere sampling, Spanish features *partido* (‘game’ or ‘(political) party’), *gato* (‘cat’ or ‘(hydraulic) jack’), and *presa* (‘prey’ or ‘dam’). A small offering of the many cases of synonymia in English, each with at least three meanings, includes ‘date’ (a day on the calendar, a romantic outing, or a fruit; Spanish: *fecha, cita, dátil*), ‘party’ (a social gathering, a group of people seated together at a restaurant, or a political organization; Spanish: *fiesta, grupo, partido*), and ‘spring’ (a season of the year, a metal coil, or a place where water emerges from the earth; Spanish: *primavera, resorte, manantial*). Yet another example in English (Poibeau, 2017: 171), processed into Spanish, will again demonstrate the possible pitfalls related to such lexemes:

(17) ‘Little John was looking for
his toy box. Finally, he found it. The box was in the **pen**. John was very happy.’

GT (9 Oct 2021):

El pequeño John estaba buscando su caja de juguetes. Finalmente, lo encontró. La caja estaba en el ***bolígrafo****. John estaba muy feliz.*

The GT rendition of example (17) is illogical to the point of being essentially impossible. The Spanish term *bolígrafo* is a common one for ‘pen’ when it denotes a writing instrument (specifically a ‘ballpoint pen’). Since it is not feasible in any reasonable way for a box of toys to fit inside a ballpoint pen, the ‘pen’ in question surely refers to a child’s playpen (*corralito*), or a pen in which animals are perhaps kept (*corral*), or a similar area of confinement. The obvious problem is that GT, which translates at the sentence level, neither has enough context to know what type of ‘pen’ might be involved nor realizes that what it has proposed is a physical impossibility. This is because a computer system trained to detect patterns does only that and does not “realize” anything the way humans do; it attempts natural language processing without the aid of “natural” (logically intuitive) capacity. Barring some unlikely paradigm shift in this regard, the type of mistake made by GT in this example seems rather insuperable.

The treatment of homonymia above serves as a segue into a distinct but related phenomenon: polysemy, in which a single word can express partially related, even somewhat overlapping concepts, each of which may have a separate term in another language. For instance, when one uses the English verb 'to save,' which in general means to keep something or someone from being harmed or lost, its different specific connotations and appropriate verb translations into Spanish include, to note a few examples, 'to save a living thing' (*salvar*) 'to save money' (*ahorrar*), or 'to save a computer file' or 'to put something away for safekeeping' (*guardar*). Without context, GT is incapable of knowing which verb to use when translating from English to Spanish. Speaking of an amount of money (*ahorrar*), or a piece of pizza (*guardar*), one might exclaim: 'I want to save it,' which GT renders as *Quiero salvarlo* (as if the speaker wanted to save a puppy). If however, one states, 'I want to save this piece of pizza,' a correct translation is given: *Quiero guardar este trozo de pizza* (10 Jun 2022).

While the cases just discussed were created as test samples by the author, the following example, which also involves polysemy, comes from transcribed dialogue in the Spanish sitcom *Aquí no hay quien viva* (Miramón Mendi, 2003). Speaking of the need to use the stairs to ascend to the apartment units above, as the elevator is old and only supposed to be used for descending, one of the actors states:

(18) *...solo lo utilizamos para **bajar**. Se **estropea** mucho.*

GT (2 Oct 2021):

'…we only use it **to download**. It **spoils** a lot.'

Author's translation:

'…we only use it to **go down**. It **breaks down** a lot.'

While the speaker did mention the elevator, it was in an earlier sentence, leaving GT, which works at the sentence level, to guess at the intended meaning. In Spanish, the verb ***bajar*** means not only '**to descend**,' but also to '**download**' (a computer file, for instance). Likewise, ***estropearse*** (perhaps employed more frequently in Spain than the common Latin American equivalent *dañarse*) can refer to something – organic or inorganic – being damaged, but it is general enough that English requires different specific verbs in translation in order to capture the precise meaning, depending on the context. English speakers may well say that a head of lettuce '**spoils**,' but not an elevator, which '**breaks**' or '**breaks down**.' If one manipulates the original sentences, intentionally joining them and repeating the word *ascensor* ('elevator') explicitly – which of course the proficient human translator does not require – then a correct, idiomatic sentence results:

Solo utilizamos el ascensor para bajar porque se estropea mucho.

GT (26 Mar 2022):

'We only use the elevator to go down because it breaks down a lot.'

Two final sets of examples will be given, one in which Spanish differs from English (and perhaps other languages), and the other in which it is English that features the marked construction and is prone to causing erroneous GT renditions. The first example, taken from Bolivia's *El Espectador* newspaper, concerns time-related references in Spanish:

(19) *Los tres países afinan detalles para firmar el acuerdo que pondrá en marcha el proyecto para verificar la destrucción de cocales, anunciado **en marzo pasado**.*

GT (1 Apr 2022):

The three countries are fine-tuning details to sign the agreement that will launch the project to verify the destruction of coca crops, announced **last March**.

The otherwise impressive GT rendition into English only becomes problematic at the end of the sentence. The article in question is dated 19 Apr 2011, and the March that is mentioned is the previous month, literally the 'last' one to transpire, which is precisely how *pasado* is used in Spanish. In English, however, the correct translation is 'in March of this year,' or, in this specific case, 'last month' would also suffice. This same issue of temporal orientation can exist when a Spanish-language text refers to a future date, as seen in the following text from Argentina's *Cronista* newspaper:

(20) *El* ***próximo jueves*** *24 de marzo se recuerda a las víctimas de la última dictadura con el Día Nacional de la Memoria por la Verdad y la Justicia.*

GT (1 Apr 2022):

Next Thursday, March 24, the victims of the last dictatorship are remembered with the National Day of Memory for Truth and Justice.

The article in question appeared on Tuesday (22 Mar 2022), which means that the 'next Thursday' would indeed technically be on day 24 of the month. However, the date in question would be accurately expressed in English as 'this Thursday,' or 'Thursday of this week,' since 'next Thursday' would be used to designate Thursday, 31 March, a full week later.

However, it is English that at times poses a unique challenge to GT in its use of the modal verb 'should' to convey not the semantic conditional, but rather the habitual past, such as in this following mini-dialogue of the author's creation:

(21) Person 1: 'What **would you do** on the weekends?'

Person 2: 'We **would go** to the beach.'

GT (5 Mar 2022):

Person 1: *¿Qué* ***harías*** *los fines de semana?*

Person 2: ***Iríamos*** *a la playa.*

Regarding the translation for Person 1, such a phrase in Spanish would only be used literally, such as in the following sentence expressing a hypothetical: *¿Qué* ***harías*** *los fines de semana* ***si tuvieras*** *más tiempo y dinero?* 'What **would you do** on the weekends **if you had** more time and money?' Of course there is more than one way to express many if not most ideas. For instance, each person in the dialogue could have recast their part thus: 'What did you **use to do** on the weekends?'; 'We **used to go** to the beach.' Just as with the original phrases, Spanish uses the imperfect aspect of the past tense to accomplish this – *hacías* and *íbamos* or *solías hacer* and *solíamos ir* – never the conditional. A lack of context, however, can cause GT to opt for a literal translation of 'should,' changing the intended meaning entirely.

4 Conclusions

This paper is part of a larger, long-term study whose central focus is the ability (or inability) of Google Translate (GT) to render acceptable translations among multiple written genres between English and Spanish and vice versa. Some of the challenges relating to this pair of languages extend to others. While many GT capabilities have been greatly enhanced since the service's 2016 shift from the use of statistical machine translation (SMT) to a system of neural machine translation (NMT), this does not mean that all such renditions are perfect or even acceptable, or that its performance based on the perceived complexity or simplicity of source texts is predictable. Some economic texts, for instance, are rather intricate, but GT more often than not produces very usable English-Spanish-English results. In contrast, some seemingly simple texts are badly distorted when run through GT, even when no ambiguity is readily apparent. This is surely a reflection of the fact that even apparently uncomplicated human language is more involved than its speakers typically realize. Added to this is the fact that even relatively simple ideas can be expressed in such a variety of ways that no database could contain all the possibilities, let alone adequate past translations of them. This means that any solution to mistranslations could either be years, even decades away, or, surely in some cases, never be attainable at all, signifying that human translators will need to continue occupying an indispensable role in the translation process for the foreseeable future. While ascertaining some of these matters to the degree possible is the objective of the larger study, various patterns have already begun to suggest themselves and have been demonstrated to a modest degree in this paper via several examples featuring Spanish, English, and Bulgarian.

References

El Cronista. 22 Mar 2022. 'Próximo Feriado del 24 de marzo: ¿es puente, hay fin de semana largo?' Retrieved 15 March 2022.

El Espectador (EFE). 19 April 2011. 'Bolivia acepta ayuda económica de EE.UU. para destruir la coca.' Retrieved 30 March 2022.

'Érase una mudanza.' *Aquí no hay quien viva*. Season 1, episode 1, Miramón Mendi S.L., 2003.

Fitzgerald, Maggie and Stevens, Pippa. 10 Aug 2021. 'Dow rises 220 points to new record after inflation report is not as bad as feared.' *cnbc.com*. Retrieved 4 Oct 2021.

Garavalova, Iliyana. 2020. 'The Bulgarian dialect names of the cat.' *Papers of BAS. Humanities and Social Sciences*, 7(2), pp. 103-116.

Koehn, Philipp. 2020. *Neural Machine Translation*. Cambridge, UK: Cambridge University Press.

Leafgren, John. 2011. *A Concise Bulgarian Grammar*. Durham, NC: Slavic and Eurasian Language Resource Center.

Lewis-Kraus, Gideon. 14 Dec 2016. 'The Great A.I. Awakening.' *The New York Times*.

Piqué, Elisabetta. 2014. *Francisco: vida y revolución: Una biografía de Jorge Bergoglio*. Chicago: Loyola Press.

Poibeau, Thierry. 2017. *Machine Translation*. Cambridge, MA: MIT Press.