WHERE GOOGLE FAILS

Maria Stambolieva

New Bulgarian University, Sofia, Bulgaria

Abstract

The paper presents ongoing research in contrastive corpus linguistics with envisaged applications in machine translation (MT) and with focus on Google Translate (GT) performance in English-Bulgarian translation. Structural patterns, forms or expressions where automatic translation fails are identified and analysed in view of creating a GT-editing tool providing improved target language output. The paper presents the corpus and the corpus analysis method applied, including the identification of inacceptable string types, their structural analysis and categorization. For each failure type, pre- or post-GT editing transformations are proposed. A first outline is proposed of a GT-editing tool consisting of a pre-GT editor performing string identification, substitution or deletion operations, a post-GT editor with a set of more complex string transformation rules and an additional module transferring structural information.

Key words: machine translation, pre-editing, post-editing, Google Translate, bitext, computational linguistics, corpus linguistics

Article history:
Received: 8 November 2015;
Reviewed: 20 November 2015;
Revised: 29 November2015;
Accepted: 30 November 2015;
Published: 31 December 2015

Maria Stambolieva, PhD in Linguistics from St. Kliment Ohridski University of Sofia, Bulgaria, is Associate Professor at the English Studies Department, New Bulgarian University, Sofia, and head of the NBU Laboratory for Language Technologies. M. Stambolieva participated in a large number of European, bilateral and large national projects in the field of computational linguistics, corpus linguistics, formal and contrastive linguistics and the application of computational corpus linguistics in (specialised) foreign language teaching. She is the author of over 70 publications.

Email: mstambolieva@nbu.bg

Approaches to Machine Translation

In computational linguistics, translation is viewed as a particular type of paraphrase of a text written in Language A (LA, the source language), where the paraphrase is a text written in Language B (LB, the target language) and where "paraphrase" is defined as "a restatement of a text or passage giving the meaning in another form" (Cf. e.g. Teubert, 1997, p. 147).

The first attempts in the field were inspired by computer-assisted code breaking during the Second World War, the hypothesis being that machine translation is a task comparable to deciphering coded messages. Translation at this stage was *direct*, i.e. LB → LB – lexical substitutions without intermediate representations, and *local* – mainly simple reordering transformations in the immediate neighbourhood of a unit.

Probably the best example of direct translation is the Georgetown automatic translation system (GAT). The project started in the early 50es and the system was operational from 1964. GAT was used for the translation of specialised texts in the field of physics, from Russian into (something resembling) English. In the early 60es, GAT was replaced by its modern version – SYSTRAN, which latter was used in Google translation tools until 2007.

The development of formal grammars (initially mainly of the transformational type) allowed the transition from local to *global* approaches to translation, with representations and cross-language transformations on the level of the clause and sentence. Expectations, however, were unreasonably high; the 60es ended with general disappointment and the notorious ALPAC report which brought this early period of research to an end. In the early 70es, only three MT projects received government funding in the USA; not a single project in the field was funded in 1975. Even so, many government agencies continued to use MT systems of the 60s – simply because they had no alternative for the purpose of rapid translation.

The 80es saw a renewed interest in machine translation, this time with more modest output expectations and with the understanding that good MT performance does not necessarily exclude human intervention. Systems are evaluated in terms of two criteria:

1. whether the output quality is good enough to serve as raw text for human editing and
2. whether the development is justified in terms of price, speed or other factors (Cf. e.g. Slocum, 1985).

Approaches to MT in the 50es and 90es can be illustrated with two quotations from leading representatives in the field, Warren Weaver and John R. Searle. Although not the first to imagine automated translation, Warren Weaver won himself the reputation of father of both machine translation and computational linguistics with the ideas he put forward in a 1947 letter to Norbert Wiener:

> "I have a text in front of me which is written in Russian, but I am going to pretend that it is really written in English and that it has been coded in some strange symbols. All I need to do is strip off the code in order to retrieve the information contained in the text." (Arnold et al., 1994, p. 13).

John Searle's no less famous Chinese Room Argument is a good illustration of late 20th century developments in the field:

> "Imagine that I, a non-Chinese speaker, am locked in a room with a lot of Chinese symbols and boxes. I am given an instruction book in English for matching Chinese symbols with other Chinese symbols and for giving back bunches of Chinese symbols in response to bunches of Chinese symbols put into the room through a small window. (…) the symbols put in through the window are called questions. The symbols I give back are called answers (...). The boxes of symbols I have are called a database, and the instruction book in English is called a program. The people who give me questions and designed the instruction book are called the programmers, and I am called the computer. We imagine that I get so good at writing the program, that eventually my 'answers' to the 'questions' are indistinguishable from those of a native Chinese speaker. I pass the Turing test for understanding Chinese. But all the same, I don't understand a word of Chinese and – this is the point of the parable – if I don't understand Chinese on the basis of implementing the program for understanding Chinese, then neither does a digital computer solely on that basis because no digital computer has anything that I do not have." (Searle, 1995, p. 546)

The major obstacle to the success of rule-based translation tasks is the first part– the formulation by a human /humans of a working set of instructions, of the "instruction book" type – because a set of working instructions would include: an algorithm for correct parsing and generation at the levels of morphology, syntax, semantics and pragmatics and taking into consideration context, world knowledge and the culture of the speaker/hearer. It would also require the resolution of different types of ambiguity, name recognition, anaphora resolution, in-depth understanding of the meaning of the message. (Cf. Nakov, 2012, pp. 110-111). Hence the turn towards computer assisted translation (workbenches of the TRADOS type), statistical and example-based translation. The most prominent example of these latter is the modern, post-SYSTRAN, version of Google Translate.

Google Translate

Google's translation tool is a development initiated by Franz Josef Och (head of Google's machine translation group until 2014) – a switch from SYSTRAN (still used by Yahoo! Babel Fish) to example-based, statistical machine translation. This is how Och explained the switch:

> "Most state-of-the-art commercial machine translation systems in use today have been developed using a rule-based approach and require a lot of work by linguists to define vocabularies and grammars. Several research systems, including ours, take a different approach: we feed the computer with billions of words of text, both monolingual text in the target language, and aligned text consisting of examples of human translations between the languages. We then apply statistical learning techniques to build a translation model." (Quigley, 2010).

Och criticised the effectiveness of rule-based[1] algorithms[2], with their increasing complexity, and argued in favour of statistical approaches. Extrapolating from language teaching, we could say that Google switched from a language learning to a language acquisition approach to translation. In this approach, the larger the amount of text sifted, the more successful the result will be, as Och himself remarked:

> "…Once the computer finds a pattern, it can use this pattern to translate similar texts in the future. When you repeat this process billions of times you end up with billions of patterns and one very smart computer program. For some languages however we have fewer translated documents available and therefore fewer patterns that our software has detected. This is why our translation quality will vary by language and language pair." (Quigley, 2010).

A sound basis for a working MT system, according to Och, would be a bilingual corpus of at least 150 million words – and the more, the better. Not surprisingly, GT performs remarkably well for the six official UN languages and for languages with longer EU status, while providing much poorer results for Bulgarian.

[1] https://en.wikipedia.org/wiki/Rule-based_machine_translation

[2] https://en.wikipedia.org/wiki/Algorithm

All automatic translation tools have their limitations, and GT is no exception. Some of the serious problems the GT research team themselves admit are: the (lack of) analysis of the category of Mood and, ever more dramatic, the difficulty to cope with the expression of habituality and progress, as marked by Romance perfect and imperfect tenses.[3] The failure zones in the English-Bulgarian corpus used are more numerous, ranging from incorrect analysis of the source strings, generation of ungrammatical target strings, or both, and covering the levels of the lexicon, of morphology, syntax and semantics. The aim of the research is to determine the feasibility of achieving improved machine translation output by enhancing GT performance with an automatic pre- and post-GT editing tool. I view of this task, inacceptable strings are first identified, following which structural types are systematized and editing transformations are proposed.

The Corpus

The corpus used for the research is based on a Teach Yourself manual of English consisting of short graded dialogues in English and their Bulgarian translations (Stambolieva, 2016). It consists of 250 dialogues at levels L1 (Starter-A2), or approx. 4, 500 sentences, and 100 dialogues at level L2 (B1), approx. 5, 500 sentences.

For each of the two levels, the dialogues (D) are grouped into Units, with three subparts each, each dialogue set consisting of 10-15 paired sentences for the first level and 15-20 paired sentences for the second level, where each dialogue is provided with a (human) translation, aligned with the source at sentence level.

The bitext is turned into a tritext by aligning the human translations (HT) with GT-generated translations:

L(evel)1. 1-1-D(ialgue)3.

Source	HT	GT
1. What’s this?	Какво е това?	Какво е това?
2. It’s an Ili pika.	Това е Или пика.	Това е Ili пика
3. Do you like it?	Харесва ли ти?	Харесва ли ти?
4. I do.	Да.	Аз правя.

[3] https://en.wikipedia.org/wiki/Google_Translate#Limitations

5. It’s cute.	Сладко е.	Това е сладък.
6. Is it a mouse?	Мишка ли е?	Дали това е мишка?
7. It isn’t.	Не е.	Не е.
8. Does it look like a mouse?	На мишка ли прилича?	Изглежда ли като мишка?
9. The body, yes.	Тялото да.	Тялото, да.
10. The face is like a teddy bear’s.	Лицето е като на мече.	Лицето е като мече е.
11. But the ears aren’t.	Но ушите не са.	Но ушите не са.
12. They are like a rabbit’s.	Те са като на заек.	Те са като на заек.
13. It’s a very rare animal.	То е много рядко животно.	Това е много рядко животно.
14. Where does it live?	Къде живее?	Къде се живее?

Aligned sentences with clear HT-GT asymmetry are marked (as are the underlined 4, 5, 6, 8, 10, 13, 14,above). GTs which are acceptable translations (8, 13, possibly 6 too) are treated as legitimate paraphrases of the HTs.

At this first stage of research a subcorpus of 30 dialogues at Beginner level are analysed for the purpose of identifying instances where Google fails and the structural types of failure. At a second stage of the research, the incorrectly generated target structures will be paired with transformation types.

Google Translate Fail Areas

Two major phenomena, looming large in natural language, lead to inaccurate translation: ambiguity and multi-word expressions. Ambiguity can be observed on both the lexical and the grammatical (above all morphological) level. In rule-based systems, where translation is preceded by source text analysis, ambiguity resolution is based on the analysis of context. The presentation which follows is a first step towards the study of typical problem areas in English-Bulgarian example-based translation; further, it presents a basis for the formulation of rules directed toward improved target string acceptability.

The Lexicon

Some multiple-word expressions, such as idioms or compounds, are listed in the lexicon of translation systems as separate items. Natural language further abounds in a number of set phrases, but also grammatical structures, for which word-for-word analysis is out of place. Complex lexemes can be compound nouns, syntactic fragments, even sentences. Additional difficulties stem from the fact that the building blocks of these lexemes need not be adjacent or otherwise fixed. Ivan Sag (Sag et al. 2002) identifies three categories of complex lexemes: a. fixed: *in short*, *by and large*, *every which way*; b. semi-fixed – with possible morphological variation: *car park/parks, kick/kicked the bucket*; c. flexible: phrasal verbs and light verb constructions: *pick* [someone] *up*, *make a mistake*, etc. Most flexible structures also allow for syntactic transformations, such as passivisation.

In statistical translation, problems are not simply related to identifying multiple-word expressions but also to resolving ambiguity between word combinations, collocations and different types of multiple-word expressions. These problems are well exemplified in the corpus in the following translation triples:

15. Be off to (I'm off to school) – GT: На разстояние съм от (училище)! HT: Отивам на училище.
16. Take off that hat! – GT: Излитане, както шапка! HT: Свали си шапката!
17. How do you do. – GT: Как го правиш? HT: Приятно ми е.
18. How about your sister? – GT: Какво ще кажеш за сестра ти? HT: Ами сестра ти?
19. Here's your change. – GT: Ето ти промяна. HT: Заповядайте рестото.
20. Good night – GT: Добър вечер. HT: Лека нощ.
21. I've got a date. – GT: Имам една дата. HT: Отивам на среща.
22. He is six. – GT: Той е шест. HT: Той е на шест.
23. You look your age. – Изглеждаш Вашата възраст.

The GT output in the above examples is rather surprising, as the translations offered reflect neither greater frequency of occurrence of the proposed sense of the input string in English (as can be observed in the British National Corpus), nor any probability of occurrence of the Bulgarian output string worth registering. A pre-GT editor for the MWEs should contain, in the simplest case, lists of expressions, with translation equivalents. The pre-GT editor will substitute target language strings for the source language ones:

17a. [How do you do, Mr. Jones. $_{\text{source string}}$] → [Как сте, Mr. Jones. $_{\text{pre-editor output}}$] → [Как сте, г-н Джоунс. $_{\text{GT output}}$]

A slightly more sophisticated pre-editor, with context sensitive rules and access to a semantically annotated lexicon, would output „е на шест" for the source string "is six" of 22. above in the context of a preceding [+Human] NP – pronoun, noun or anthroponym.

The Morphological Level

Aspect. Translators from Bulgarian to English and back, and their editors, point to Aspect as a major pitfall. Aspect is, again, the category where systems for automatic translation seem to offer the least help – Cf. the translation equivalents provided by Google Translate for a few English sentences:

24. He sang the song. → Той изпя песента. (Perfective Aspect, Aorist)

25. He sang for an hour. → ?То пееше за един час. (Imperfective Aspect, Imperfect Tense)

26. They ate the sandwich. → *Те яде сандвич. (Imperfective Aspect, Aorist/Present?)

27. Did you eat the sandwich? → *Знаете ли, яде сандвич? (???)

The problem with the translation of Aspect in English-Bulgarian translation is that while Bulgarian aspect is an equipollent lexico-grammatical category covering the entire verbal system and unambiguously defined in the lexicon (the semantic basis of the opposition being the presence or absence of a bound ([+Bound] / [-Bound]) in the topological structure of a situation), few – if any – of the defining features of the Slavonic category can be said to be applicable to the English data. The grammatical system of English does incorporate an opposition of an aspectual type – the so-called "Progressive Aspect'. This, however, is a privative opposition between an unmarked form and a marked form expressing non-boundedness, plus a large number of other components of meaning, of non-topological nature – such as limited duration, irritation and other nuances of emotional colouring, increasing or decreasing activity, etc. The non-progressive form in the English "aspectual" opposition is unmarked with respect to boundedness. In other words, the English non-progressive verb cannot unambiguously define a situation as eventive or not. Seeing that, on average, English non-progressive forms occur approximately 20 times oftener than progressive ones in an English narrative text, this means that *English verbs are, largely, unmarked for boundedness.*

Henk Verkuyl (1972, 1993 and following publications) demonstrated that in non-Aspect languages such as English, events are construed, i.e. boundedness obtains at VP and sentence level as a result of the combination of verbs belonging to particular verb classes with quantified or unquantified complement or subject NPs. About the same time and independently of Verkuyl, Danchev, & Alexieva (1974) in their English-Serbo-Croatian and English-Bulgarian contrastive studies, respectively, arrived at similar results, namely: aspect markers in English occupy a large stretch of the discourse. While Ridjanovic (1969) concentrated on the articled/non-articled noun phrases as major markers of Aspect, Danchev & Alexieva, processing a large parallel corpus (20 000 file-cards of English Simple Past Tense sentences and their Bulgarian equivalents) arrived at a much greater variety of contextual markers. The authors ranked these as follows: adverbial phrases, verb semantics, subject phrase semantics, object quantification.

The analysis of the corpus points to the following major Perfective Aspect contextual markers in the English sentences:

Adverbial modifiers of time:

- *when* - upon concordancing, found to present, in about all cases, an instance of the relative adverbial, introducing a time clause;

- *then*, *now*, *now that*, *before*, *as* (=when), *eventually*, *finally*, *in+year* (e.g. *in 1984*), *at lunch*, *to begin with*, *the moment +subject+V*.

Coordination:

Coordinative links between event clauses: conjunctions and commas.

Lexical meaning of the verbs:

- communication verbs in the simple past tense, esp. *admitted, announced, insisted, lied, mumbled, prompted, said, thought (to myself), urged*;

- phrasal verbs: *drove away*, *went away*, *sat down*, etc.

- process verbs in the simple past tense.

The following were found to be the major Imperfective Aspect markers in the corpus:

Adverbial modifiers:

- temporal adverbials, e.g. *still*, *sometimes*, *repeatedly*, *when* (=*whenever*, closely followed by *would*), *as* (= *while*)

- *for*-phrases: e.g. *for a few minutes*;

- *do nothing but*, e.g. *We did nothing but quarrel*.

- adverbial modifiers of time containing NPs with attributes pointing to iterative situations, e.g. *every day*, *every summer*.

The lexical meaning of the verb:

- link verbs, e.g. *was*, *seemed*, *grew*;

- extended state verbs, e.g. *know*, *hope*, *love*, *remember*;

Subject phrase semantics:

- Subjects semantically characterised as [-Animate], and esp. 'Inalienable property' subjects, e.g. *the symmetrical limbs*, *her expression*, etc. are typical clauses with Imperfect Aspect readings.

The major contextual markers of aspect were systematised in Stambolieva (2012). Most of them can be integrated into a GT editing tool consisting of a pre-translation editor and a post-editor, with information transfer from the former to the latter.

Grammatical Homonymy. One of the major sources of NLP difficulty, or failure, is homonymy. Grammatical homonymy is a particularly important obstacle to the automatic processing of English.

The following cases of analysis failure in the bicorpus are due to incorrect resolution of grammatical homonymy in the source language:

1/ *‘s* is a possessive case marker, but can also be a contracted form of the 3rd p. sg. form of the auxiliary or link verb *to be*. *That* could be either a demonstrative pronoun or a sentential conjunction (complementizer). The following GT outputs demonstrate that 1.

the same values are by default attributed to all occurrences of the forms and 2. very low probability strings appear in the Bulgarian output:

26. Excuse me, that's my magazine. –> GT: Извинете ме, че това е моето списание. HT: Извинете, това е моето списание.
27. Peter, please give me that pen. –> GT: Моля да ми дадете, че писалка. HT: Питър, подайте ми писалката, ако обичате.
28. It's my father's, actually –> GT: Това е баща ми е, всъщност. HT: Всъщност е на баща ми.

In a rule-based approach, the analysis of the input string would be unproblematic. In 27: *that* can only be a determiner, since a complementizer is followed by a clause. In 28: the contracted form of the auxiliary or link verb must be followed by either a non-finite verbal form or a noun phrase/adjectival phrase/adverbial phrase/ prepositional phrase; further, it is not followed by a comma. In this occurrence, *'s* can only be a marker for the Genitive and can very easily be detected as such, and translated, by a pre-processor.

2/ Definitions and examples of *to be* as an existential full verb can be found almost exclusively in dictionaries. In actual text, the forms of this verb are instances of either the auxiliary or the link verb, and a non-contracted form followed by a punctuation mark normally occurs in short answers only. Because such short answers are not present in the structure of Bulgarian, the best target string in such cases is, simply, *Да*.

29. Are you Bulgarian? – I am. → GT: (…) – Аз съм. HT: - Да.

Search for these strings and the substitution of short positive or negative answers with *да* or *не*, respectively, can be another pre-processing task.

2/ English *to do* occurs most frequently as an auxiliary in questions, negative forms, short answers and emphatic structures. There is also a semi-auxiliary (or light verb) and a full verb. Only the latter could be translated with the Bulgarian full verbs *правя*. It is therefore surprising to find *правя* as the exclusive translation equivalent of *do* in the corpus – Cf. example 30 below (where the translation of *do* is not the only problem!):

30. What does one wall say to the other? → GT: Какво прави една от стените се каже на другия? HT: Какво казва едната стена на другата?

The translation can be improved by pre-GT *do*-deletion, post-GT editing of the Bulgarian text, or – best – both. *Do*-deletion gives a slightly improved output, which is easier to edit:

30a. What one wall say to the other? → Какво едната стена се каже на другия?

3/ *One* is, of course, a numeral, but also a pronoun. The ambiguity of the string does not seem to have hit the processor, which is odd: a simple BNC search demonstrated that the probability of its occurrence as a pronoun is, if not greater, at least the same as that of the numeral.

31. -- A blue one or a green one? – A green one→ GT: -- А синьо един или зелена един? – А зелен един. HT: Синя или зелена? -- Зелена.

In any rule-based system, the above string would be unambiguously analysed as a NP. Numerals are not heads of phrases containing determiners and attribute adjectives. A simple rule stating that if *one* appears at the end of a phrase, it is most probably a pronoun, would be sufficient to resolve the ambiguity of the input string. This, paired with the information that *един/една* etc. are seldom pronouns in Bulgarian, would do away with the second word in the output string. A least effort solution, in this case too, would be (*one*-) deletion at the pre-GT editing stage plus simple post-GT editing:

31a. A blue or a green? → А син или зелен?

4/ Bulgarian has both full pronouns and short pronominal forms (clitics). There are possessive clitic pronouns for all three persons and numbers. Only one form appears in the output offered by GT, however: the short reflexive pronoun *си*. It is difficult to believe that these results are example-based:

32. Where are your wives? → GT: Къде са жените си? HT: Къде са жените Ви?
33. These are your keys. → Това са ключовете си. HT: Това са ключовете Ви.
34. There's a fly in my soup. → Има една муха в супата си. HT: Има муха в супата ми.

The translation can be improved with a rule-based Bulgarian post-editor.

5/ The comment for the clitic examples fully applies to the translations of English *it*, which in all its manifestations – as substitute for [-Human] nouns, as an impersonal subject or a dummy subject, is invariably translated with the Bulgarian demonstrative pronoun *това*:

35. It is not there. → GT: Това не е там. HT: Не е там.
36. It's here. → GT: Това е тук. HT: Тук е.
37. It's on the fourth floor. → GT: Това е на четвъртия етаж. HT: На четвъртия етаж е.
38. It is in the hotel. → GT: Това е в хотела. HT: В хотела е.
39. It is rainy and cloudy. → GT: Това е облачно и дъждовно. HT: Облачно и дъждовно е.

The simple transformational rule applied by the human translator is based on the fact that, first, Bulgarian does not have dummy or impersonal subjects (hence the *It* position is erased), second, that a Bulgarian sentence cannot begin with a clitic form (including verbal forms) (hence the link verb is removed from its position in the English sentence) and third, that sentences are not made up of string of words but of phrases (and therefore the verbal clitic does not appear after the first word but after the first phrase). In the above cases, post-GT editing would be best, for *It*-deletion does not result in improved GT-output:

35a. Is not there → Не е ли там.

39a. Is rainy and cloudy. → Дали дъждовно и облачно.

6/ English adjectives and nouns for nationalities coincide in form, which is not the case in Bulgarian. The following Bulgarian string from the bi-corpus was clearly not directly drawn from a corpus of parallel text. It is based on a higher frequency of occurrence of the adjectives, disregarding context.

40. Are you Italian? → GT: Италианската ли сте? HT: Италианка ли сте?

The context-sensitive rule here would be that *Italian* in the immediate context of the personal pronouns *I, you, he, she* or *we* is to be translated in Bulgarian as *италианец / италианка.* Sticking to our proposal for a GT+ analysis, post-GT editing of the Bulgarian text would be easiest.

7/ Finally, the following GT outputs are, I must admit, difficult to account for:

41. Don’t you have a bag? → GT: Не трябва да имате една торба. HT: Нямаш ли чанта?
42. You haven’t got a driving license? → GT: Вие не сте ли шофьорска книжка? HT: Нямате шофьорска книжка?

Ungrammaticality of the target string. Along with the numerous instances of ungrammaticality due to the incorrect analysis of the input string, several cases of unacceptable output strings are the result of a combination of insufficient corpus length and lack of rules:

1/ The plural forms of Bulgarian masculine nouns change under quantification – Cf. *много сандвичИ* (many sandwiches) – *няколко сандвичА* (several sandwiches). This rule is simple enough to formulate, but requires POS tagging and analysis. Lacking that, GT yields strings like 45. and 46. below.

43. Fifteen sandwiches → GT: Петнайсет сандвичи. HT: Петнайсет сандвича.
44. How many sandwiches? → GT: Колко сандвичи? HT: Колко сандвича?

2/ Definiteness receives only one marker in the Bulgarian noun phrase – and in this respect the two languages are similar. GT correctly outputs possessive pronouns marked for definiteness (even though English possessive pronouns are inherently definite and do not need the marker) but nevertheless outputs ungrammatical structures with double marking of the category.

45. our room keys → GT: нашите ключовете. HT: нашите ключове.
46. (I give him) his coat and hat → GT: неговото палтото и шапката. HT: палтото и шапката (му)/ неговото палто и шапка.

3/ Bulgarian, unlike English, has Vocative case. Nouns marked for the Vocative appear in clear syntactic positions, usually at the head of the phrase, and punctuation sets them apart from the rest of the clause. Not a single Vocative was presented as output by GT for our corpus.

47. - Yes, mother. → GT: - Да, майка. HT: - Да, майко.

Note that the three cases of ungrammaticality presented in this section do not lead to unintelligibility and can be resolved with a set of relatively simple post-GT editing rules.

Poor Syntax

1/ Lack of agreement. Achieving correct Gender/Number agreement with an example-based translator would require a very large and varied corpus. Clearly, such a corpus is not yet available in Google, which results in the generation of the following types of ungrammatical output strings:

a. Lack of agreement within the Noun Phrase:

48. What a nice piano!→ GT: Каква хубава пиано! HT: Какво хубаво пиано!
49. What instrument do you play? → GT: Какво инструмент да играеш? HT: На какъв инструмент свириш?

b. Lack of agreement between the nominal part of the predicate and the subject:

50. It's not very hot. → GT: Това не е много горещ. HT: Не е много горещо.
51. It's cheap. → GT: Това е евтин. HT: Евтино е.
52. It's elegant but rather expensive. → GT: Това е елегантен, но доста скъпи. HT: Елегантна е, но доста скъпа.
53. She is cute. → GT: Тя е сладък. HT: Тя е сладка.
54. That bag is heavy. → GT: Тази чанта е тежък. HT: Тази чанта е тежка.
55. Because the world is round.→ GT: Защото светът е кръгла. HT: Защото светът е кръгъл.

c. Lack of agreement between the verb-predicate and the subject:

56. They may have a phone. → GT: Може би те има телефон. HT: Те може да имат телефон.

d. Between the main clause (modal verb) and a subordinate:

57. You can easily get lost here. → GT: Можете лесно да се загубиш тук. HT: Можете лесно да се изгубите тук.
58. Can you show me the way? → GT: Можеш ли да ми покаже пътя? HT: Можете ли да ми покажете пътя?
59. He can't call her. → GT: Той не може да ѝ се обадя. HT: Той не може да ѝ се обади.

The ungrammaticality problems of 43-59 can be resolved with the help of a post-GT editor.

2/ Poor syntax of interrogative sentences.

a. Tag questions

Tag questions are quite easy to identify. For them, the simple rule should be to substitute all tags, positive and negative alike, with *нали*? at the pre-GT stage. This simple operation could considerably improve on GT outputs:

60. (It's rather cold,) isn't it? → GT: не е тя? → HT: нали?

b. Yes/No questions

For Yes/No questions, we have identified three types of ungrammatical outputs.

Inappropriate *дали*-insertion and missing interrogative particle *ли* :

61. Is she a new student? → GT: Дали тя нов ученик? HT: Тя нова ученичка ли е?
62. Are these your parents? → GT: Дали тези Вашите родители? HT: Това родителите Ви ли са?

Note that, for some reason, the GT output containing *дали* does not include finite verbal forms. The transformation necessary in these instances will include *дали*-deletion and the insertion of *ли*, followed by a form of *to be* after the NPs, or auxiliary-deletion at the pre-GT stage, followed by simpler post-editing.

The interrogative particle is, again, missing in cases where *дали* is not inserted – which is odd, seeing that it is the major signal for interrogation in Bulgarian grammar:

63. Are schools in England like Bulgarian schools? → GT: Има училища в Англия като българските училища? HT: Училищата в Англия като българските ли са?
64. Is it far from here? → GT: Далеч от тук е? HT: Далече ли е оттук?

Clearly, one of the important functions of the post-editor would be the restructuring of strings signalled by a question mark.

Incorrect analysis of *that*:

Finally, the systematic interpretation of English *that* as a complementizer results in the following output strings (which, along with being ungrammatical, are also unintelligible):

65. Is that your father? → GT: Е, че баща ти? HT: Това баща ти ли е?
66. Is that the sun or the moon? → GT: Е, че слънцето или луната? HT: Това слънцето ли е или луната?

67. Is that your daughter, Mr. N? → GT: Е, че дъщеря Ви? HT: Това дъщеря Ви ли е, г-н Н.?

A post-GT rule of sentence-initial „Е, че" substitution with „Това" plus final „ли е" insertion could be a possible solution.

3/ Non-causatives analysed as causatives.

The structural asymmetry between English and Bulgarian is very marked in the area of diathesis, but it is nevertheless surprising that a default ergative reading should surface in the translation.

68. She doesn't cook. → GT: Тя не се готви. HT: Тя не готви.

In this case, more complex automatic editing would be necessary, involving both a pre-GT analysis of the immediate context (to eliminate the ergative reading) and the transfer of this information to the post-editor.

Conclusions

The alignment of the English-Bulgarian bicorpus with the output of Google Translate demonstrated that, while Google is beyond doubt a very useful tool, it can also output inaccurate, ungrammatical, occasionally even unintelligible target strings. The analysis of this negative output allowed the identification of several substructures where analysis, generation, or both, systematically fail. Main lexical and structural problem types were identified and editing procedures were proposed. A first outline was proposed for a GT-editing tool consisting of a pre-GT editor performing string identification, substitution or deletion operations, a post-GT editor with a set of more complex string transformation rules and an information transfer module between the two.

References

Arnold D.J., Balkan, L., Meijer, S., Lee Humphreys, R., & Sadler, L. (1994). *Machine Translation: An Introductory Guide*. Manchester/Oxford: NCC/Blackwell.

Danchev, A., & Alexieva, B. (1974). Izborat mezhdu minalo svarsheno i minalo nesvarsheno vreme pri prevoda na the Past Simple Tense ot anglijski na balgarski ezik. [The choice between the Aorist and the Imperfect in the translation of the Past Simple Tense from English to Bulgarian]. *Faculty of Classical and Modern Languages Yearbook, LXVII*(1), 249-329.

Desclés, J.-P. (1990). *State, Event, Process and Topology*. *General Linguistics, 29(3),* 159-200.

Nakov, P. (2012). Savremenen statisticheski mashinen prevod. [Modern statistical machine translation]. In M. Stambolieva (Ed.), *Kompyutarna lingvistika 1, Problemi i perspektivi [Computer Linguistics 1, Problems and Perspectives]* pp. 110-155. Sofia: ANABELA. Retrieved from http://people.ischool.berkeley.edu/~nakov/selected_papers_list/nakov_prevod_sp_Avtomatika_Informatika.pdf

Quigley, R. (2010). *How does Google Translate work*. Retrieved from http://www.themarysue.com/how-does-google-translate-work

Ridjanovic, M. (1969). *A Synchronic Study of Verbal Aspect in English and Serbo-Croatian*. (Doctoral dissertation). University of Michigan.

Sag, I., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002). Multiword Expressions: A Pain in the Neck for NLP. *Lecture Notes in Computer Science, 2276*, 1-15.

Searle, J. (1995). John R. Searle. In: S. Guttenplan (Ed.), *A Companion to the Philosophy of Mind* (pp. 544-550). Basil Blackwell Ltd., Oxford,

Slocum, J. (1985). A Survey of Machine Translation: Its History, Current Status and Future Prospects. *Computational Linguistics*, *11*(1), 1-17.

Stambolieva, M. (2008). *Building Up Aspect: A study of aspect and related categories in Bulgarian, with parallels in English and French*. Peter Lang: Oxford, Bern, New York.

Stambolieva, M. (2012,). Parallel Corpora in Aspectual Studies of Non-Aspect Languages. *Proceedings of the Second Workshop on Annotation and Exploitation of Parallel Corpora, RANLP 2011*, 39-42. Shoumen: Incoma Ltd. Retrieved from http://www.aclweb.org/anthology/W11-4306

Stambolieva, M. (2016). *Angliyski ezik – Samouchitel I [English language – Self study I]*. GRAMMA Publishers: Pleven (in print).

Teubert, W. (1997). Translation and the Corpus. In R. Marcinkeviciene, & Volz, N. (Eds.), *Proceedings of the Second TELRI Seminar on Language Applications for a Multilingual Europe*, 147-164. Kaunas: Lithuania.

Verkuyl, H. (1972). *On the Compositional Nature of the Aspects*. Reidel: Dordrecht.

Verkuyl, H. (1993). *A Theory of Aspectuality: The interaction between temporal and atemporal structure*. (Cambridge Studies in Linguistics). Cambridge:Cambridge University Press.