

## The Translation Game

FEDERICO GOBBO

*University of Insubria Varese-Como, Italy*

### ABSTRACT

*Evaluation of Machine Translation (MT) quality is an open issue among specialists, and a serious philosophical investigation is still needed. MT output evaluation can become a site of corpus-based testing of the appropriateness of the models of the language faculty currently available, providing the evaluation is performed in a setting that minimizes the effect of prevalent negative attitudes against MT. This paper proposes a Gedankenexperiment called the Translation Game. It is designed to enable an evaluation of MT output from a maximally neutral standpoint. Certain objections are articulated and addressed.*

### 1. INTRODUCTION

The evaluation of machine translation (MT) quality is an open issue among specialists, and a serious philosophical investigation is still needed. What exactly is involved in “machine translation?” Let us initially consider translation as a process of rendering an asynchronous written text, i.e. a coherent chain of grammatical sentences written in a given natural language (NL), from a source language ( $L_s$ ) into a reliable text written in a target language ( $L_t$ ). By the term “reliable” we mean that (a) the text in  $L_t$  is syntactically well-formed and does not call for any further editing; (b) the original meaning in the  $L_s$  is preserved in the  $L_t$  version, i.e., both texts convey the same content. This definition of reliability is less strong than the notion of authenticity drawn from translation studies by Bachman-Palmer (2000) in that we do not identify any explicit task/s in the target language that can be used to measure the quality of translation. That criterion of “authenticity” severely restricts the range of texts for which one can procedurally validate a translation; for this reason, I choose to adopt the notion of reliability as a basis for the present discussion.

By MT I mean translation carried out by a computer. This means that the translation input is performed by a human being, in the format suitable for the particular MT system (this phase is known as “pre-editing” in MT parlance) while the translation process is performed automatically, without any human aid. The procedure counts as MT proper if its output is regarded as the final translation – if, in other words, no “post-editing” is performed by humans. When the procedure does include a post-editing component, it becomes Computer-Aided Translation (CAT). Normally MT is performed with texts written in an objective language register – such as papers in political science – where the content is specifiable at the level of truth value. In texts where the meaning is highly subjective, i.e. dependent on the author, and the value of the text crucially invokes ambiguity and the individual identity of the author as factors that the reader must take into account – such as poetry or literary prose – MT is far from being effective. In this paper, the focus is on texts that can be handled by an MT system.

Why should such cognitive science disciplines as linguistics, computational linguistics or the philosophy of mind be concerned with the evaluation of MT output? MT output evaluation can be considered a corpus-based test of the appropriateness of the models of the language faculty available, provided that the evaluation is performed in an appropriate setting, where translationese – i.e., the set of linguistic indicators that identify the text as a translation – is judged without psychological prejudice. Unless a specific context is constructed to avoid it, such prejudice tends to be present, because human informants are perfectly aware that the text to be evaluated is a MT product and cannot eliminate this awareness when they are requested to evaluate such a text.

MT has one of the longest histories of any Artificial Intelligence enterprise. The first MT experiment was carried out by IBM at Georgetown University in 1954 (Hutchins 1997). After an initial period of high hopes, MT failed to reach its ambitious goal of building a general system able to translate every type of text without human revision. As this hope receded, the attitude of linguists, professional translators and computer scientists towards MT became more and more negative. Furthermore, in recent years some elementary MT systems such as Babelfish have become freely available to the general public through the internet: since these systems are not specialized, the quality of their output is highly dependent on the register of the input text, and in most cases the results are grossly inaccurate. For all these reasons, there is a general bias against MT, which needs to be taken into account if an objective evaluation of MT results is our goal.

The community of MT researchers tried to solve this problem by introducing automatic metrics that would make it possible to compare the performance of two or more MT systems. However, in the last ten years or so, this practice has been questioned by some influential members of the community itself. While automatic metrics – such as BLUE, trained on a gold standard reference corpus from the  $L_t$  – are suitable tools for the MT engine development phase, they cannot replace human beings as the final evaluators (Callison-Burch, Osborne & Koehn 2006). In general, human evaluation is subject to great variability and is not easy to control or measure. Given the psychological bias toward MT explained above, whenever human evaluators know that an output has been produced by a machine and not by human translators, this knowledge deeply influences, or even invalidates, their final evaluation.

To address this important issue, I propose here an alternative scenario, where the evaluators do not know in advance that they will be evaluating an MT-produced text. I shall call this scenario the Translation Game – a name that recalls the Imitation Game (Turing 1950), one of the key moves in the early work that established Artificial Intelligence as a viable research programme. The Translation Game is a *Gedankenexperiment* – a thought experiment along the lines of Turing (1950) and Searle (1980). This scenario, I will argue, makes it possible to make at least a partial advance over earlier forms of MT evaluation.

## 2. THE TRANSLATION GAME SCENARIO

Let us consider first the default scenario of MT. Suppose that Alice is a Spanish native speaker and she wants a document of hers, e.g. a newspaper article, to be read by Bob, a native speaker of Tamil who does not understand Spanish. It is not easy to find professional translators from Spanish to Tamil; so Alice decides to invest in MT, as her article will be the first of a series. Suppose that Charles is the Spanish-Tamil MT designer. Alice learns how to type her Spanish text in the software user interface given to her by Charles. Bob will read the MT output, without being in any direct contact with Alice.

Now we modify this default scenario by introducing a slight but important change. Charles asks Alice not to write directly in her mother tongue but in a special controlled language, by which we mean not a domain restricted sublanguage but a full language system of the type that has been articulated in terms of the concept of a Quasi-Natural Language (QNL). Lyons (2006) characterizes a QNL as being non-natural but not an “unnatural” language in the sense in which the

predicate calculus or computer programming languages depart from the naturalness of human language. In particular, a QNL shares all the design features of a NL, the most important being the double articulation in terms of phonemes and morphemes. As a direct corollary, QNLs are not restricted in semantics, i.e., they have the same expressive power as NLs. Furthermore, a QNL has the following properties:

- a) it is highly regular in morphology;
- b) as a direct corollary, homophony and suppletion as lexical strategies are avoided or eliminated;
- c) compared with NLs, part-of-speech tagging is easier as there is little or no allomorphy;
- d) lexeme formation strategies are much more productive than in NLs.

QNLs are not semantically restricted in semantics, but have the same expressive power as NLs. It is noteworthy that both non-pathological child languages and planned languages belong to this category.

For example, QNL has a subclass of Quasi-Englishes, one of whose members is like English in all respects except that it is inflectionally regular, all plurals of nouns being formed with the *-s* suffix (“childs,” “sheeps,” “gooses” etc.), all past-tense forms of verbs with *-ed* (“goed,” “runned,” “beed” etc.), and so on. Children construct part of this language on their own (and then in part dismantle) at a certain stage in the “natural” process of acquiring English. This Quasi-English is the form that the NL English would presumably have taken if it had developed under particular environmental conditions maximizing the effect of what is traditionally referred to as analogy (Lyons 2006: 69-70).

Let us call the QNL used by Alice in the Translation Game the “Entry Language” (EL), which in the approximation under discussion at this point is some version of Quasi-Spanish. The EL is parsed by the MT engine that generates the translations in Spanish and in Tamil at the same time. The EL should simplify not only Charles’ work, but also Alice’s. The system not only renders Alice’s EL text simultaneously in Spanish and in Tamil, but also enables her to compare her text with the Spanish MT output and hence adjust her EL text. She is thus able to learn how to improve her text and her use of the MT system simultaneously. In particular, the system maintains a Translation Memory (TM), which proposes second or third choice alternative translations. In fact, when Alice judges the Spanish MT to be satisfactory, she pushes an “ok, I’m satisfied” button. From that

moment onwards, the system keeps track of Spanish and Tamil texts at the TM level – developing the capacity to modify its generation rules on the basis of the versions judged to be satisfactory, and thus to come up with better proposals for the Spanish output. In the background, this exercise keeps modifying the Tamil generated text too, along with the generation algorithms responsible for Tamil. The Translation Memory database is a structured collection of examples; the reference is to “machine translation by example-guided inference, or machine translation by the analogy principle” (Nagao 1984: 4).

Note that from Bob’s point of view, nothing has changed. He may be totally unaware of all these processes; all he needs is to be sure about reliable Tamil translations – in the sense or *reliability* specified above – of Alice’s Spanish originals. The main advantage of the Translation Game scenario lies in the MT evaluation test we are going to perform, a test that is not possible in the default scenario.

### 3. A NEW EVALUATION TEST

We now introduce a human evaluator; call him Dave. Let us assume that Dave is (a) a bilingual native speaker of Spanish and Tamil, (b) knows neither Alice nor Bob, (c) is neither a computer scientist nor a linguist, but a professional translator, (d) is unaware of the existence of the EL, and (e) has not been told that MT is involved.

Dave receives from Charles a text in Spanish and a text in Tamil and is asked to answer the following question: “Are the two texts reliable translations of each other? To what extent?” Recall assumption (e): Dave has not been told that MT is involved. The point of designing the Translation Game is that it should be possible to obtain a neutral evaluation, and it should be impossible for Dave to guess that a machine is involved in one direction or the other, because under these assumptions it is a machine, not a human, that has produced *both* the texts. Notice too that in this design the question by Charles does not specify the translation arrow – Dave has no way to know which is the source language.

Suppose that Charles’ MT system has finished its fine-tuning procedures and has thus acquired a satisfactory TM, thanks to judgments provided by Alice. Under such optimal conditions, Dave should be unable to decide whether the Spanish text was the original version or not. At the level of establishing goals, I propose that the system be regarded as passing the test if Dave finds the same amount of translationese –linguistic markers in a document that indicate that it is a translation – in both the texts. These proposals are in no way

inconsistent with the use of any automatic metric; the test suggested here is simply an additional test with design features that distinguish it from human evaluation tests based on the default MT scenario.

#### 4. ESPERANTO AS THE ENTRY LANGUAGE

When we imagine Alice providing her input in Quasi-Spanish, we are trying to work with a minimal modification of the default scenario. This model is of some heuristic use as an expository step as we build up the Translation Game. But a moment's reflection shows us that designing a pedagogy for Alice that would enable her to construct inputs in Quasi-Spanish rather than the NL Spanish that she is used to, or formulating a full and explicit characterization of Quasi-Spanish (and Quasi-NL<sup>1</sup> for any other language) and feeding this characterization into the MT system to enable the system to deal with Alice's input, is by no means a straightforward task. In other words, if we wish to use the Translation Game scenario schema as a basis for models that can lead to actual experimentation, it makes sense to consider more than minimal modifications of the default scenario.

Any EL with the design features required for the purposes of such an experiment will have to be a specifically constructed language. One way to save effort is to choose a candidate for the EL that has already been constructed and for which explicit descriptions, suitable for MT, exist. The obvious choice would appear to be Esperanto. It is the most widespread constructed language; it is a QNL in the sense of the earlier discussion; and MT experience has shown that its design features make it suitable as an intermediate language for MT use. For readers unfamiliar with the field of constructed languages, these points need some elaboration.

From the final decades of the 19th century up to about 1950, the optimal design for a definitive international auxiliary language (IAL) was a topic of serious debates among academic linguists and other stakeholders. In fact, about 1,000 language projects were under active consideration in that period; most of these proposals had originated in Europe. In terms of our discussion here, we may regard an IAL as a QNL designed for a specific purpose, for "oral and written use between people who cannot make themselves understood by means of their mother tongues", to return to a lucid statement of this purpose by Otto Jespersen from 1931. IALs, often called planned languages, are complete linguistic systems – *langues* in the Saussurean sense of this term – launched by an author through a book (or, in more recent times, through a web site) that provides a grammar and a basic vocabulary. At

that moment, an IAL is a langue without any *parole* – the question of the linguistic behaviour of members of its speech community does not arise until there is such a community. Note that this definition, since it envisages the possibility of a speech community, excludes such “a priori languages” as pasigraphies, or John Wilkins’ *Real Character*, or François Sudre’s *Solresol*.

Esperanto was launched by Ludwik Lejzer Zamenhof (1859-1917). He developed an inter-ethnic bridge language with the goal of enabling persons from any culture to use this neutral language and transcend difficulties arising from an exclusive focus on particular ethnicities or nationalities. In 1887 he launched the language by publishing a book in Russian that provided the grammar, some literary texts (original and translated) and the basic vocabulary. Translations into Polish, French, German, English and Swedish followed. Zamenhof published his work under the pseudonym ‘Doktoro *Esperanto*’ – *hopeful person* – which the rapidly growing community of users adopted as the name of this new IAL.

These users were not simply individuals corresponding with the author of the project. Local clubs and national associations focused on the idea of an IAL – relatively numerous and visible at the end of the 19th century – lent their support to Esperanto and organized its users into a worldwide network. This unusual speech community demonstrated its resilience by surviving two World Wars and direct targeting by Hitler and Stalin (Lins 1988). Esperanto has been the only constructed language to have emerged as the carrier of a non-ethnic international culture including a serious body of original and translated literature (Sutton 2008) – which is of interest in the present context since MT systems need corpus data for their training.

The use of Esperanto in MT has a relatively long history. The Soviet scientist Petr Petrovich Troyanskii, who has been called “the Babbage of machine translation,” acquired a patent in 1933 for a mechanical translating apparatus (it comprised a desk, a typewriter, a camera and a belt); he drew on the Esperanto repertory for his symbols of logical and etymological parsing. It was the Stalin regime’s suppression of Esperanto that led Troyanskii to stop using it (Hutchins-Lovtskii 2000). Long before the recognized take-off of MT in the 1950s, then, Esperanto was a factor in the field.

MT depends for its implementations on a formal description of the languages involved, and usually syntax plays an important role in the system. One of the pioneers in the domain of formal syntax, Lucien Tesnière – whose work appeared posthumously (Tesnière 1959) and is widely used, by scholars in linguistics and language teaching in Central

and Eastern Europe – drew on Esperanto for its basic symbolic devices as well. Among the MT systems that have used Esperanto, the Distributed Language Translation (DLT) project is worth mentioning because it was based on a Tesnièrean dependency grammar (Schubert 1986, 1987, Maxwell & Schubert 1989). DLT, which started with a seminal study by A. P. M. Witkam in 1982, was developed over ten years. A prototype was presented in 1987 and led to a commercial version in 1993. These versions had English and French as source and target languages; Esperanto was the basis of the interlingua module. This project demonstrated the viability of Esperanto as an interlingua for MT use. For an independent evaluation, see Hutchins & Somers (1992: ch. 17).

#### 5. ADDRESSING CRITICISM

The tradition initiated by such classic papers as Turing (1950) and Searle (1980) makes it appropriate to summarize some possible contrary views to the approach presented, along with replies to these objections.

##### *The Chinese Room argument*

“Whatever linguistic model you put into the machine, we cannot consider the procedures as constituting real cognition, for the meaning of the linguistic model exists only in the brain of Charles, the system designer.” This argument is based on Searle (1980). I think that MT is one of the best testing grounds for explicitly stated linguistic models. Even if the linguistic model is meaningful only in Charles’ brain, the model becomes explicit at the level of the computer program. Furthermore, the model is refined through tests that are part of the computer implementation.

##### *The engineer’s reaction*

“Machine translation is not a testing ground for linguistic theories or anything else. What we need is something practical, i.e. commercially useful in domains where translation of large amounts of data need to be delivered quickly.” This argument is seldom presented openly but reflects beliefs that are widely held. My reply is that the decision to adopt no linguistic theory also amounts to a linguistic model and needs to be compared with other linguistic models based on a particular theory. Close inspection of even the field of statistical MT shows the need to resort to syntax when one seeks to improve the quality of translations. Computational brute force and statistics are not enough for

achieving good results in MT (Callison-Burch, Osborne & Koehn 2006).

*The typologist's objection*

“Esperanto is fundamentally a European-based language; your scenario may work with English, French or Spanish, but not with non-European languages, such as Chinese, Arabic, or Tamil.” This argument is potentially serious; I have only a partial answer to offer. Esperanto uses morphemes that are essentially pan-European, but its morphology surprisingly resembles such non-Indo-European language systems as for instance Hungarian or Turkish (Gledhill 2001). It is possible that the Translation Game gives rise to greater difficulties when languages remote from Indo-European are involved (and forces “Alice” to work harder if she is a speaker of, say, Tamil or Turkish). This is an empirical question. Note that the issue does not pertain to the structure of the Translation Game scenario, but carries over the properties of the default MT scenario. If Spanish to Tamil MT is intrinsically harder than Spanish to English MT, it follows that the Translation Game scenario will replicate this relative difficulty.

*The human interface argument*

“Your assumptions are too strong. You are forcing Alice not to use her mother tongue, i.e., Spanish, and you are asking her also to learn the system that Charles has set up. Furthermore, your approach involves strong supervision. Surely it will be easier, faster and cheaper to have professionals translate from the source language to the target language than to use any variant of your system.” This pragmatic argument is couched in economic terms. However, the claim that Alice is not allowed to use Spanish misrepresents the set-up. She is made to choose between alternative Spanish versions of each syntagm that the system presents to her as part of the procedure. As noted in the description of the scenario, Alice’s additional work makes it possible to arrive at a rigorous characterization of the relation between the EL and the source language; in economic terms, this characterization should be accepted as an outcome making Alice’s labour a valuable contribution to the development of the system. Turning to the strong supervision issue, such supervision is required only in the analysis phase, and involves only a monolingual parser – the generation of natural language string pairs does not involve metataxis in the sense of Tesnière (1959). As the TM grows, the system becomes more precise and its coverage expands. This feature of the procedure can be strengthened by a decision to release such an MT system as an open source project, and thus to draw

on the verbal and scientific resources available in the worldwide Esperanto speech community, whose members all speak two or more languages and often translate.

#### 6. CONCLUSION

Subjective bias can affect the quality of MT evaluation; the Translation Game proposes a new way of addressing this issue. Moving from this proposal into actual implementations will obviously depend also on concrete matters outside the purview of the present exercise. However, the design of the Translation Game makes it possible to connect specific issues in the field of MT – including issues pertaining to the validity of certain set of methodological assumptions often adopted in such inquiry – to linguistic theories and to specific characterizations of human cognition and behaviour in this domain.

#### NOTE

1. For readers unfamiliar with the background from which this project originated, a brief characterization may help. Zamenhof was an Ashkenazi intellectual who participated in the Jewish Enlightenment (Haskalah) and early versions of Zionism. Taking the debate about the possible solutions of the Jewish question as a point of departure, Zamenhof spent his life developing a twofold project: a ‘neutral human’ language and a ‘neutral-human’ culture and nonclerical religion. While his neutral religion project did not take off, his linguistic project, Esperanto, became the best known constructed language of all time.

#### REFERENCES

- Bachman, L. F. & Palmer, A. 2000. *Language Testing in Practice*. Oxford: Oxford University Press.
- Callison-Burch, C. & Osborne, M. & Koehn, P. 2006. *Re-evaluating the Role of BLUE in Machine Translation Research*. In Diana McCarthy & Shuly Wintner, (Eds.), *Proceedings of EACL-2006. 11th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 249-256), University of Trento, Trento, Italy.
- Gledhill, C. 2001. *The Grammar of Esperanto. A Corpus-based Description*. Munchen: Lincom Europa.
- Gobbo, F. 2005. The European union’s need for an international auxiliary language. *Journal of Universal Language*, 6, 1-28.
- Hutchins, J. N. 1997. From first conception to first demonstration: The nascent years of machine translation, 1947-1954. *Machine Translation*, 12, 195-252.

- . & Lovtskii, E. 2000. Petr Petrovich Troyanskii (1894-1950): A forgotten pioneer of mechanical translation. *Machine Translation*, 15, 187-221.
- . & Somers, H. L. 1992. *An Introduction to Machine Translation*. London: Academic Press.
- Lins, U. 1988. *Die gefährliche Sprache. Die Verfolgung der Esperantisten unter Hitler und Stalin*. Gerlingen: Bleicher.
- Lyons, J. 2006. *Natural Language and Universal Grammar*. Cambridge University Press.
- Maxwell, D. & Schubert, K. 1989. *Metataxis in Practice: Dependency Syntax for Multilingual Machine Translation*. Dordrecht: Foris.
- Nagao, M. 1984. A framework of a mechanical translation between Japanese and English by analogy principle. In A. Elithorn and R. Banerji (Eds.), *Artificial and Human Intelligence*. Ch. 11. Elsevier Science Publishers.
- Schubert, K. 1986. *Syntactic Tree Structures in DLT*. Dordrecht: Foris.
- . 1987. *Metataxis: Contrastive Dependency Syntax for Multilingual Machine Translation*. Dordrecht: Foris.
- Searle, J. 1980. Minds, brains and programs. *Behavioral and Brain Sciences*, 3/3, 417-457.
- Tesnière, L. 1959. *Eléments de syntaxe structurale*. Paris: Klincksieck.
- Turing, A. 1950. Computer machinery and intelligence. *Mind*, 59, 433-460.

**FEDERICO GOBBO**

DIPARTIMENTO DI INFORMATICA E COMUNICAZIONE,  
UNIVERSITY OF INSUBRIA VARESE-COMO, ITALY.  
E-MAIL: <FEDERICO.GOBO@UNINSUBRIA.IT>