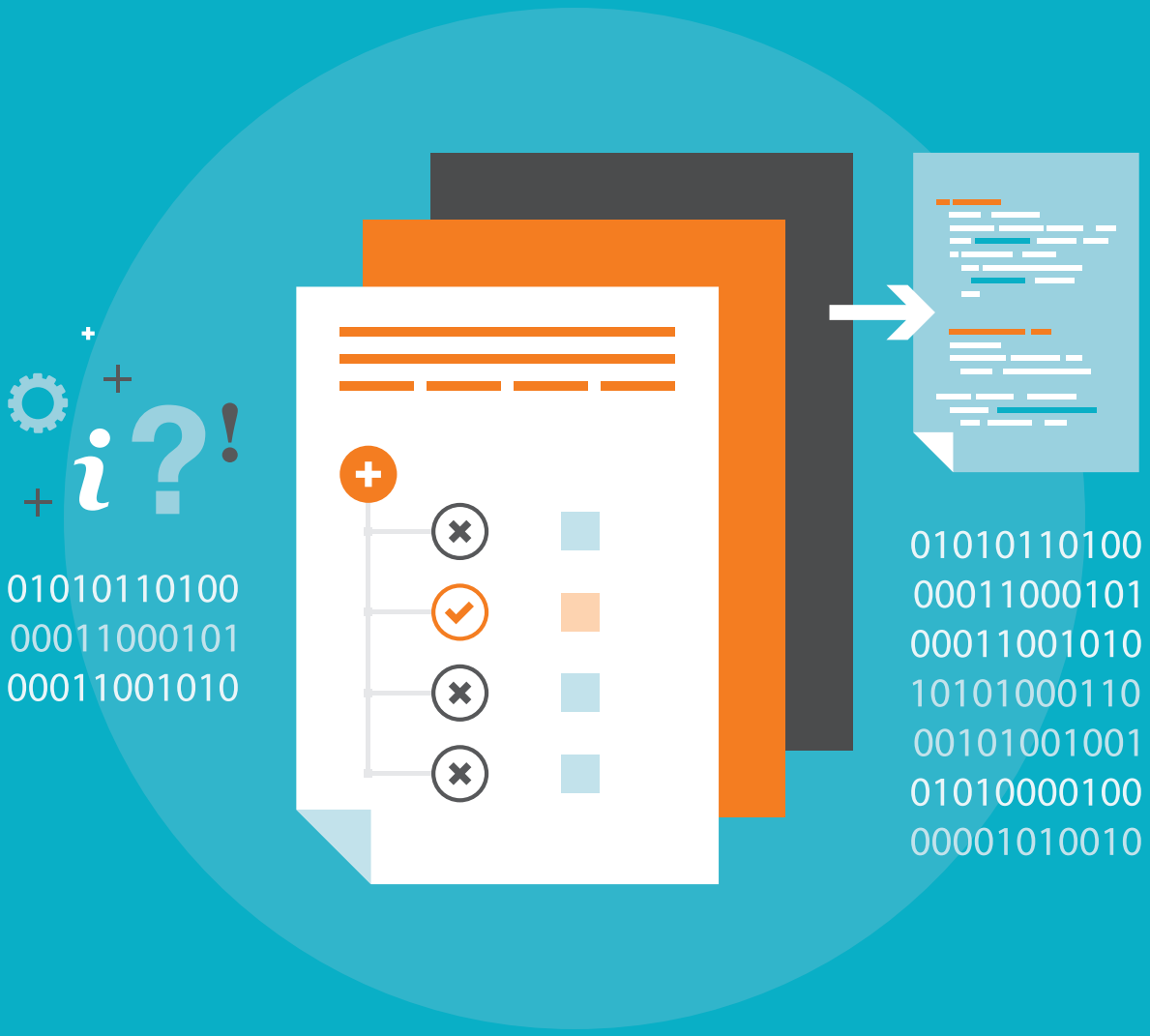
TQ-AUTOTEST: NOVEL ANALYTICAL QUALITY MEASURE CONFIRMS THAT DEEPL IS BETTER THAN GOOGLE TRANSLATE

By Vivien Macketanz, Aljoscha Burchardt & Hans Uszkoreit

GALA
Globalization & Localization Association



TQ-AutoTest: Novel analytical quality measure confirms that DeepL is better than Google Translate

By Vivien Macketanz, Aljoscha Burchardt & Hans Uszkoreit

In the second half of 2017, a surprising news was that a small German company has built DeepL, a Machine Translation system that was reported to be able to beat Google and other known systems in terms of translation quality (see, e.g., https://techcrunch.com/2017/08/29/deepl-schools-other-online-translators-with-clevermachine-learning/). Using our new semi-automatic tool TQ-AutoTest that allows for an informative, analytical comparison of different MT engines, we could confirm this observation for the language pair German – English that we have examined.

Introduction

Assessing translation quality is notoriously difficult. In the area of Machine Translation (MT) research and subsequent marketing, simple automatic comparisons of system output and human reference translations are usually taken as approximate indications of quality. The automatic measures used are BLEU, Meteor, and others. They have major shortcomings in that they do not provide reliable assessment of single sentences, that they do not provide any indication about the nature and severity of the error, or that they do not allow for a meaningful comparison across tools, languages, etc. Sometimes, A-B tests involving humans are used to compare systems' quality, but they also do not provide any insights about the particular strengths and weaknesses of the systems.

In a sequel of EC-funded projects (QTLaunchPad, QTLeap, QT21) we have devised a new method for assessing MT Quality in close cooperation with GALA and GALA members (LSPs, translators, researchers). Our method can be classified as a source-driven approach as opposed to the prevalent reference-based paradigm. The basic idea is to use a suite of segments exhibiting relevant (linguistic) phenomena and to assess the systems' performance on each phenomenon individually. We understand linguistic phenomena in a broad way ranging from punctuation to very specific ones such as preposition stranding. Testing is semi-automatic, supported by the tool TQ-AutoTest as described below. The result is a quantitative and qualitative insight into the systems' performance such as "system X gets 20% of the negations right, and 70% of the lexical ambiguities".

In this white paper, we will briefly describe the test suite and TQ-AutoTest tool and then showcase its application in a comparison of DeepL and Google.

The test suite approach

In a team of translators and other language experts, we have built a test suite for a fine-grained evaluation of MT quality for the language pair German – English. In brief, it contains segments selected from various bilingual corpora and drawn from other sources such as grammatical resources, e.g., the TSNLP Grammar Test Suite (Lehmann, Sabine, et al. "Tsnlp: Test suites for natural language processing." *Proceedings of the 16th conference on Computational linguistics-Volume 2.* Association for Computational Linguistics, 1996.) and online lists of typical translation errors. In contrast to a "real corpus", in our test suite, segments have been shortened and made up to separate phenomena.

Test suite layout

Each test sentence is annotated with a phenomenon category and the phenomenon it represents. An example showing these fields can be seen in Table 1 with the first column containing the source segment and the second and third column containing the phenomenon category and the phenomenon, respectively. The fourth column shows an example machine translation and the last column contains a minimal post-edit of the MT output. The examples only serve illustrative purposes.

Source	Category	Phenomenon	Example Target (raw)	Target (edited)
Lena machte sich früh vom Acker	MWE	Idiom	Lena [left the field early].	Lena left early.
Lisa hat Lasagne gemacht, sie ist schon im Ofen.	Non-verbal agreement	Coreference	Lisa has made lasagne, [she] ist already in the oven.	Lisa has made lasagna, it is already in the oven.
Ich habe der Frau das Buch gegeben.	Verb tense/ aspect/mood	Ditransitive - perfect	I [have the women of the Book].	I have given the woman the book.

Table 1: Example test suite entries German → English (simplified for display porposes).

In our latest version of the test suite, we have a collection of about 5,000 segments per language direction that are classified in about 15 high-level categories (most of them similar in both language directions) and about 120 specific phenomena (many of them similar across language directions).

Depending on the nature of the phenomenon, each is represented by at least 20 test segments in order to guarantee for a balanced test set. The categories cover a wide range of different grammatical aspects that might or might not lead to translation difficulties for a particular MT system (or a human translator).

Automation of testing

Until recently, we have checked systems' output manually, which meant a lot of effort even though we focus the evaluation on the respective phenomenon under consideration. For example, if we test negation, we will ignore how the system translates, e.g., the verb or noun.

With the test suite growing bigger over time, we decided to implement a framework that facilitates the evaluation procedure by automating the analysis. Therefore, we built the TQ-AutoTest. In order to include as many correct translations options as possible, the TQ-AutoTest is based on regular expressions, i.e., sequences of letters and "wild cards".

With the help of these patterns, we try to automatically cover as many correct translations as possible. We did not only create positive regular expressions with which the MT output can be evaluated as correct, but in some cases also negative regular expressions with which the MT output is detected as incorrect. A screenshot of a positive match with the positive and negative regular expression in the TQ- AuteTest can be seen in Figure 1.

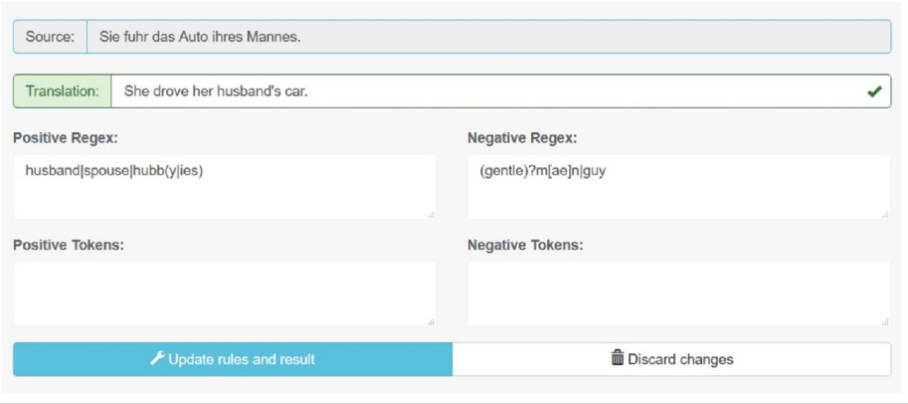


Figure 1: Screenshot of positive match and regular expressions.

The German source sentence contains a lexical ambiguity: The German word "Mann" can either mean man or husband. In combination with a possessive pronoun (in this case "ihr" - her), "Mann" always refers to husband. This is captured in the positive and negative regular expressions. If a given system output matches any of these terms, the result will be counted as positive or negative. If no regular expressions match, it is saved for manual inspection.

Comparison of systems

For the experiment presented here, we have analyzed Google Translate (based on a neural network), DeepL translator (also based on a neural network) and Lucy (based on a rule-based system.

The result of the combined automatic and manual analysis is first of all a table that indicates the percentage values of correct translations for the categories or phenomena. To get a more detailed insight and directly compare the different translations, the TQ-AutoTest also provides an overview of the translations of the different systems.

Quantitative comparison

The three systems that we analyzed are based on two different machine translation models: While Google and DeepL use a neural network, Lucy is a rule-based translation system that has been implemented with linguistic rules of the languages. Since the two approaches are very different, you can also expect the performance of the systems to be different.

The quantitative result of the analysis with the percentage of correct translations per category can be found in table 2 below.

Each of our linguistic categories contains a different number of fine-grained linguistic phenomena. For this reason, the number of segments per category (seen in the second column of the table) varies. Especially the category Verb tense/aspect/mood has a much bigger number of test segments than the other categories which is why we will look at this category separately.

Our key finding is that even though the Google and DeepL percentages are quite similar, our test suite approach confirms that DeepL performs better than Google in all categories except one (subordination). The fact that

	# items	Google (NMT)	DeepL (NMT)	Lucy (RBMT)
Ambiguity	80	64.5%	74.4%	60.0%
False friends	36	69.4%	83.3%	63.9%
Verb valency	47	57.4%	91.5%	27.7%
Non-verbal agreement	41	90.2%	92.7%	51.2%
Subordination	91	74.7%	72.5%	33.0%
MWE	36	41.7%	66.7%	25.0%
LDD & interrogatives	172	69.2%	77.3%	51.7%
NE & terminology	84	75.0%	81.0%	70.2%
Coordination & ellipsis	80	56.3%	58.8%	32.5%
Composition	46	73.9%	95.7%	80.4%
Function words	73	63.0%	89.0%	24.7%
Verb tense/aspect/mood	4475	69.0%	71.6%	83.0%

Table 2: Percentage values of systems' performances on the categories.

DeepL's performance is better than Google's has already been found by others, e.g., Techcrunch (https://techcrunch.com/2017/08/29/deepl-schools-other-online-translators-with-clever-machine-learning/). Google performs particularly well, namely with above 90% of correct translations, on one category: non-verbal agreement. DeepL performs particularly well on 3 categories: verb valency, non-verbal agreement and composition. Verb valency is furthermore the category with the biggest gap between the performance of the two systems as DeepL achieves 34.1 percentage points more than Google.

Nevertheless, the rule-based system Lucy should not be forgotten. Even though there are categories that Lucy performs poorly compared to the neural systems, there are other categories like ambiguity, false friends, named entities (NE) & terminology, and composition in which Lucy's performance comes close the other systems. And of course, it performs best of all three system's on the large category verb tense/aspect/mood which is not surprising as verb paradigms are part of the grammatical system that Lucy is build on.

Conclusively it can be said that while Lucy performs good on the grammatical basis, DeepL and Google are more flexible when it comes to "meaning" in context. Scientifically, we would consider it worthwhile to invest in a hybrid system that is composed of a neural as well as a rule-based system in order to combine the best features from both approaches.

Qualitative comparison

In addition to the quantitative comparison which is based purely on numbers, we would like to provide a more detailed insight of some example translations of the system comparison. These examples also show that translations of the same segment provided by different systems may be very variable but nevertheless correct.

1. Lexical ambiguity

Source:	**Er hat einen Kater, weil er sehr tierlieb ist.**	
Google:	He has a cat because he is very fond of animals.	✓
DeepL:	He has a hangover because he loves animals.	✗
Lucy:	He has a tomcat because it is very animal-loving.	✓

The German source sentence in example (1) contains the ambiguous noun "Kater" which can in English either refer to a hangover or a male cat. The given context of a person loving animals ("weil er sehr tierlieb ist") disambiguates the sentence, resulting in only one possible semantic meaning of "Kater", namely cat. Thus, a correct translation would have to contain the English "cat" or "tomcat", while a translation containing the noun "hangover" would be incorrect and lead to a rather curious translation. The DeepL system is mousetrapped and translates "Kater" incorrectly as "hangover". The outputs by Google and Lucy on the other hand both clearly involve a cat and are therefore correct.

2. Internal possessor

Source:	**Die Mutter hat sie am Kopf gestreichelt.**	
Google:	The mother stroked her head.	✓
DeepL:	The mother stroked her head.	✓
Lucy:	The mother stroked it at the head.	✗

In German, external and internal possession (from the category non-verbal agreement) are realized in different syntactic structures. The distinction between external and internal objects vary between languages and English does not distinguish between them syntactically. Hence, the literal translation of a German construction with an internal possessor will lead to an incorrect output. For example in (2) a literal translation would be "stroked her at the head". None of the systems at hand produces this output but the output of Lucy is still incorrect. Google-SMT and DeepL produce an accurate translation, by correctly reformulating the possession construction to "stroked her head".

3. Modal Particle

Source:	**Bist du etwa verheiratet?**	
Google:	Are you married?	✓
DeepL:	Are you married or what?	✓
Lucy:	Are you for instance married?	✗

Modal particles (from the category function words) are a phenomenon that does exist in German but not in English which makes it difficult to translate sentences with these words. Modal particles can for example be used to express a speaker's opinion or expectation, or to refer to a common knowledge between to speakers. The modal particle "etwa" in example (3) is used in questions when a negated answer is expected (in this case whether the addressee is married). In most cases, the best solution to translate a sentence with a modal particle into English is to simply leave out the modal particle, as is the case in the Google output. A literal translation of the modal particle usually leads to either an ungrammaticality or a translation that does not match the source sentence. This case can be seen in the Lucy output: Even though the translation is correct, it does not match the source. The output by DeepL includes the colloquial "or what?" at the end of the sentence which mirrors the German colloquiality caused by the use of a modal particle.

4. Collocation

Source:	**Vor dem Essen decken wir den Tisch.**	
Google:	Before the meal we cover the table.	✗
DeepL:	We'll set the table before dinner.	✓
Lucy:	Before the food we lay the table.	✓

A common type of multi-word expressions are collocations. A collocation is a combination of words that displays an above-average occurrence in one language. Thus, collocations are language-specific and are often translated with different components to another language. The German collocation "den Tisch decken"

equals the English expression "to set/to lay the table". Since the verb "decken" can have different meanings in other contexts, a separate translation of the two components "den Tisch" + "decken" might lead to an incorrect translation. Interestingly, the Google system is the only system that mistranslates this construction while the other two systems correctly translate the collocation with one of the English equivalents.

5. Ditransitive future I

Source:	**Du wirst der Frau das Buch geben.**	
Google:	You will give the book to the woman.	✓
DeepL:	You'll give the woman the book.	✓
Lucy:	You will give the woman the book.	✓

As mentioned before, a large amount of test segments in the test suite are from the category verb tense/aspect mood. Example (5) contains a sentence with a ditransitive verb in the tense future I. Ditransitive verbs take two objects, both in German and in English. In order for the segments of the verb paradigms to count as correctly translated, the whole segment needs to be correct as these test items purely consist of the verb, its subject and its object(s). All MT outputs of the segment in the example at hand are correct. Interestingly, the three MT systems produce three different translations. This illustrates why we consider the reference-independent approach as a good measurement for MT quality: When already a simple sentence like the one in example (5) can be translated in (at least) three different ways, longer, more complex sentences or even longer texts will allow an accordingly higher number of correct translations that can not be covered by one reference translation.

Conclusions

In this post, we have introduced TQ-AutoTest, a semi-automatic framework for evaluating Machine Translation quality in a detailed, analytical way. We have illustrated it in a comparison of the Google Translate and DeepL systems for German-English where we could confirm the observation that DeepL performs a little better than Google on our test set.

When inspecting examples, we also found that DeepL by and large produces a more natural and fluent output.

More information can be found in the following scientific publication:
TQ-AutoTest – An Automated Test Suite for (Machine) Translation Quality. In: Nicoletta Calzolari; Khalid Choukri; Christopher Cieri; Thierry Declerck; Sara Goggi; Koiti Hasida; Hitoshi Isahara; Bente Maegaard; Joseph Mariani; Hélène Mazo; Asuncion Moreno; Jan Odijk; Stelios Piperidis; Takenobu Tokunaga (Hrsg.). Proceedings of the Eleventh International Conference on Language Resources and Evaluation. International Conference on Language Resources and Evaluation (LREC-2018), 11th, May 7-12, Miyazaki, Japan, European Language Resources Association (ELRA), 2018.

Contact: Vivien.Macketanz@dfki.de

About GALA

The Globalization and Localization Association (GALA) is the world's leading trade association for the language industry with over 400 member companies in more than 50 countries. As a non-profit organization, we provide resources, education, advocacy, and research for thousands of global companies. GALA's mission is to support our members and the language industry by creating communities, championing standards, sharing knowledge, and advancing technology. For more information: www.gala-global.org.