# Terminology Translation Accuracy in Phrase-Based versus Neural MT: An Evaluation for the English-Slovene Language Pair

**Špela Vintar**
University of Ljubljana, Department of Translation Studies
Aškerčeva 2, SI - 1000 Ljubljana
spela.vintar@ff.uni-lj.si

**Abstract**

For specialised texts, the accuracy and consistency of terminology is of primary importance, yet most Machine Translation systems do not employ explicit strategies to ensure term consistency on the level beyond a single sentence. We present a multifaceted evaluation and comparison of a statistical phrase-based versus neural model of Google's translation system for the English-Slovene language pair, which consists of a document-based automatic evaluation with the BLEU and NIST metrics, an automatic evaluation of term translations using an existing termbase as reference, and a human evaluation of 300 sample sentences per MT model and translation direction. Results indicate that while neural MT regularly outperforms phrase-based MT in the overall scores, the accuracy of term translations is better only for the English-Slovene language pair and not in the Slovene-English translations. In the final part of the paper we discuss typical errors encountered in the different MT outputs.

**Keywords:** MT evaluation, terminology, neural machine translation, terminology in MT

## 1. Introduction

Neural Machine Translation (NMT) is quickly becoming mainstream and has been shown to outperform statistical, mainly phrase-based, systems (SMT) across a number of features. Most of the reported evaluations so far (Machacek and Bojar 2014, Bachdanau et al. 2015) rely on automatic metrics and show consistent improvement for almost all tested language pairs. Some authors recently performed more detailed comparisons of statistical vs. neural systems using human evaluators and a more detailed error typology (Bentivogli et al. 2016, Klubička et al. 2017), or focusing on specific properties of the machine translated output such as fluency or reordering (Toral and Sánchez-Cartagena 2017). While these fine-grained evaluations bring additional evidence that NMT represents a giant leap towards more human-like translations, results obtained in some error categories, e.g. lexical or terminology errors, are not as straightforward and do not always support the NMT's claims for supremacy.

In professional translation environments, terminology research takes up to 45% of the total working time spent on translating a text, and according to a recent study by SDL[1] terminology errors amount to over 70% of all errors found in the Quality Assurance (QA) process. Post-editing guidelines developed by organisations such as TAUS[2] or SDL[3] suggest that post-editors should pay particular attention to the consistency of terminology, because nearly all state-of-the-art MT systems still produce translations on a segment-by-segment basis and thus choose terms according to local contexts instead of entire texts.

The aim of this paper is to evaluate the quality of Google Translator (GT) NMT model (Wu et al. 2016) compared to its earlier phrase-based (PBMT) model for the Slovene-English language pair and in the specialised domain of karstology. Google released the NMT model for Slovene-English in October 2017 and to our knowledge no systematic comparison of both models has been performed to date. Apart from the automatic evaluation using metrics we specifically focus on the translations of domain-specific terms, where we describe an experiment combining automatic and manual evaluation of the translation accuracy for karstology terms.

## 2. Methods and Data

### 2.1 The Karst Corpus and Termbase

For our evaluations we used a parallel corpus of scientific abstracts and articles pertaining to karstology from two international journals, Acta Carsologica and Acta Geographica Slovenica. Both of these journals publish papers with abstracts in Slovene and English, and the latter translates entire articles either into Slovene or English so that the entire journal is fully bilingual with translations in both directions.

For our experiment we use 20 parallel texts, of which 15 were abstracts and 5 entire articles. The total size of the English part of the corpus is 25,423 tokens and 18,985 tokens for Slovene. All texts were translated twice, first with the PBMT model and then with the NMT model, both provided through the GT API. We translated and evaluated in both directions, English-Slovene and Slovene-English.

It might perhaps seem futile to evaluate a general purpose MT system such as GT on a domain-specific corpus. However, we selected the domain of karstology because it is a relatively well-known field in Slovenia with a large overlap with general language. Over 45% of Slovenian landscape is karst with some of the largest and most visited tourist attractions such as the Postojna or Škocjan Caves. As a consequence, there exist numerous online sources, often bilingual, from which general MT systems such as GT might obtain their training data.

For the evaluation of term translations we rely on the quadrilingual terminology database of karst terms

[1] http://www.sdl.com/download/the-importance-of-terminology-management/71096/

[2] https://www.taus.net/knowledgebase/index.php?title=Category:Post-edit

[3] http://www.sdl.com/download/introduction-to-machine-translation-and-postediting-paradigm-shift/58317/

compiled within the QUIKK project[4]. For the Slovene-English language pair the termbase contains 81 full entries with Slovene and English single- and multiword terms, definitions and other types of information. The termbase is concept-oriented so that it may contain several expressions for the same concept. Thus, for the concept defined as *large flat surface in karst formed by erosion and corrosion* we find the English terms *karst plateau* and *karst plain*, and the Slovene terms *kraška planota*, *kraška uravnava* and *kraški ravnik*.

### 2.2 Evaluation Methods

For the automatic evaluation of overall performance we use the BLEU (Papineni et al. 2002) and NIST (Doddington 2002) metrics, the former because it is the most widely used and the latter because it has been found to correlate well with human judgements on the document level (Peterson and Przybocki 2010). Since the initial inspection of the translations with the naked eye showed considerable variation in quality, we decided to compute the metrics for each text separately to be able to observe the variation in scores.

Next we approached the evaluation of terms and their translations. For the automatic part of the evaluation we simply identify terms in the original texts using the QUIKK termbase and check whether the aligned translated segment contains the correct equivalent. Because Slovene is a language with rich morphology, both the Slovene terms and the corpus were lemmatised to facilitate matching. Still, the termbase is relatively small and in addition focuses on karst landforms, we decided to complement these results with human evaluation to assess the translations of terms other than those found in the termbase.

For the manual evaluation we first considered using the MQM framework (Lommel et al. 2014), but decided against it because our specific focus is terminology and we would thus be able to use only the error category Mistranslation, which subsumes Terminology. Instead we produced evaluation sets of 300 random term occurrences for both systems and translation directions, which were manually checked by a domain expert. Three categories were used to annotate the term translations found in machine-translated sentences:

- **Correct**, meaning that the translation of the term is the correct equivalent in the selected domain, however regardless of the agreement, case, number or other grammaticality issues,
- **False**, meaning that the word or phrase in the translation is not the appropriate equivalent in karstology. In some cases the MT system used a more generic but still accurate expression; in such cases the domain expert used common sense to decide whether it was correct or false in the given context. For example, the karstology termbase lists *precipitation* and *precipitacija* as equivalents, but the system used *padavine* in Slovene, which is synonymous to *precipitacija* and was confirmed by the domain expert as a possible translation. For multi-word terms, a partially correct translation was considered wrong.
- **Omitted**, meaning that the term from the original sentence was skipped in the translation.

[4] http://islovar.ff.uni-lj.si/karst

In the following sections we describe the results and discuss the types of errors found.

## 3. Results

### 3.1 Automatic Evaluation

The texts deal with different topics within the domain of karst and contain varying ratios of domain-specific terms, which may help explain the high variations in the BLEU and NIST scores obtained (Table 1). For English-Slovene, the average BLEU score of 18.50 for PBMT ranges from 4.55 to 36.26, and the NMT average of 22.49 shows an even higher standard deviation (8.85) with scores from 6.79 to 43.41. Looking at individual BLEU scores, NMT outperforms PBMT for 16 out of 20 texts, and 15 out of 20 according to NIST scores. The latter do not always correlate with BLEU as the highest score of 5.92 was assigned to the NMT translation of article AGS3, which received "only" 28.42 BLEU points.

| | **English-Slovene** | | | | **Slovene-English** | | | |
|---|---|---|---|---|---|---|---|---|
| | **PBMT** | | **NMT** | | **PBMT** | | **NMT** | |
| | BLEU | NIST | BLEU | NIST | BLEU | NIST | BLEU | NIST |
| AC1 | 26.26 | 4.56 | 30.72 | 4.78 | 26.1 | 4.73 | 31.12 | 5.03 |
| AC2 | 7.86 | 2.36 | 10.85 | 2.58 | 16.95 | 3.77 | 15.66 | 3.70 |
| AC3 | 16 | 2.53 | 15.04 | 2.48 | 14.23 | 3.15 | 19.77 | 3.54 |
| AC4 | 24.84 | 4.03 | 34.47 | 4.69 | 26.99 | 4.41 | 27.65 | 4.38 |
| AC5 | 4.55 | 1.56 | 6.79 | 1.51 | 6.37 | 1.72 | 8.83 | 2.04 |
| AC6 | 18.3 | 3.13 | 20.35 | 2.93 | 28.87 | 3.97 | 34.1 | 4.28 |
| AC7 | 36.26 | 4.92 | 43.41 | 5.09 | 38.14 | 5.40 | 40.93 | 5.24 |
| AC8 | 17.76 | 3.29 | 22.57 | 3.77 | 24.23 | 4.02 | 24.13 | 4.00 |
| AC9 | 15.06 | 3.43 | 31.81 | 4.30 | 21.85 | 4.21 | 35.75 | 5.06 |
| AC10 | 15.01 | 3.52 | 18.14 | 4.12 | 23.19 | 4.34 | 23.32 | 4.28 |
| AC11 | 19.6 | 3.65 | 22.54 | 3.78 | 26.12 | 4.25 | 25.97 | 4.46 |
| AC12 | 11.76 | 2.45 | 11.05 | 2.19 | 17.49 | 3.11 | 17.63 | 3.10 |
| AC13 | 8.04 | 2.09 | 11.94 | 2.47 | 16.09 | 3.33 | 11.4 | 3.15 |
| AC14 | 21.41 | 3.87 | 29.3 | 4.28 | 27.11 | 4.71 | 35.92 | 4.79 |
| AC15 | 20.96 | 3.45 | 24.08 | 3.85 | 22.93 | 4.16 | 27.25 | 4.39 |
| AGS1 | 25.77 | 5.08 | 23.89 | 4.91 | 22.6 | 5.28 | 23.24 | 5.28 |
| AGS2 | 21.69 | 4.47 | 21.3 | 4.54 | 21.71 | 4.87 | 24.98 | 4.99 |
| AGS3 | 22.02 | 5.24 | 28.42 | 5.92 | 23.11 | 4.78 | 28.11 | 4.53 |
| AGS4 | 13.49 | 3.41 | 17.21 | 3.78 | 19 | 4.85 | 23.47 | 5.08 |
| AGS5 | 23.28 | 4.75 | 25.97 | 5.13 | 27.55 | 5.76 | 29.38 | 5.76 |
| **Average** | **18.50** | **3.59** | **22.49** | **3.85** | **22.53** | **4.24** | **25.43** | **4.35** |
| St. dev. | 7.24 | 1.02 | 8.85 | 1.13 | 6.41 | 0.90 | 7.97 | 0.88 |

*Table 1: BLEU and NIST scores for the En-Sl and Sl-En language pairs*

For Slovene-English, the scores are on average slightly higher with 22.53 BLEU for PBMT and 25.43 for NMT, and a moderate improvement in the NIST score from 4.24 to 4.35 respectively. It also seems that the average quality is slightly more consistent with English as the target, as the standard deviation is lower than for En-Sl in all four sets of scores. Again, NMT achieves higher BLEU scores for 16 out of 20 texts.

### 3.2 Evaluating Term Translations

When we automatically checked for the occurrence of terms from the termbase in the original and the presence of the correct equivalent in the translated segment, the results were inconclusive (Table 2). For the English-Slovene language pair NMT appears to choose the correct equivalent slightly more often than PBMT, while the opposite direction shows a reversed picture with PBMT outperforming NMT by 30 correct translations. It should be noted however that the figures below represent term occurrences and not different terms, thus a large portion of these examples (over 300) were simply occurrences of the terms *karst* (Sl. *kras*) and *cave* (Sl. *jama*) which were for the most part translated correctly by both systems. Of course in many cases these two words occurred within a multi-word term, but if the multi-word term was not recorded in the termbase we could not automatically detect it.

| | En-Sl | | Sl-En | |
|---|---|---|---|---|
| | PBMT | NMT | PBMT | NMT |
| Terms in original | 538 | 538 | 680 | 680 |
| Correct terms in translation | 420 (78%) | 431 (80%) | 476 (70%) | 446 (65.5%) |

*Table 2: Terms and equivalents matching the termbase*

A detailed insight into the performance of both MT system versions and the types of errors they make can only be gained through human evaluation where we consider the full terminological inventory of the texts. Here, the domain expert was advised to assess not only other multi-word terms but also the translation of proper names referring to relevant places in karst (*Divača karst, Postojna Cave*) which can be especially tricky due to the rich morphology and complex capitalisation rules in Slovene (*Divaški kras, Postojnska jama*). On the other hand, grammatical errors were not to be considered, so that a correct term in the wrong case would still be marked as correct, and the overall fluency or semantic accuracy of the sentence was not part of this evaluation.

| | En-Sl | | | | Sl-En | | | |
|---|---|---|---|---|---|---|---|---|
| | PBMT | % | NMT | % | PBMT | % | NMT | % |
| Correct | 184 | 61.3 | 211 | 70.3 | 201 | 67 | 195 | 65 |
| False | 113 | 37.7 | 85 | 28.3 | 94 | 31.3 | 99 | 33 |
| Omitted | 3 | 1 | 4 | 1.3 | 5 | 1.7 | 6 | 2 |

*Table 3: Human evaluation of term translations*

Table 3 lists the results of the human evaluation of term translations in our dataset. The best performance is achieved by NMT in the English-Slovene translation direction where over 70% of the terms were translated correctly, which is a marked improvement from 61% achieved by PBMT. However, the results for the Slovene-English language pair are less conclusive with an insignificant difference between the two system variants and with NMT performing slightly lower than PBMT, which is in line with the results from the automatic evaluation.

### 3.3 A Glance at Errors

In the English-Slovene PMBT translations, the following types of errors are most common:

- untranslated term or term component (*epigenic aquifer → epigenic vodonosnik, solution runnel → raztopina runnel, hypogenic system → hypogenic sistem, paleokarst → paleokarst*)
- ambiguous term translated with the wrong sense for the domain (*spring* /as in water spring/ → *vzmet* /as in technical domains flexible metal part/, *Mlava Spring → Mlava pomladi* /spring as season of the year/, *solution* /as in water solution/ → *rešitev* /as in solution of a problem, *cave chamber → jamski zbornice* /as in chamber of commerce/)
- errors in translations of terms containing proper names (*Carpathian karst → Karpatih kras, Divača karst → Divača kras*)
- "strange" errors, such as *karst → kra* /which is a non-existent wordform in Slovene/

In the NMT translations we encounter even more examples of translations which are difficult to explain, but on the other hand NMT is creative in coining translations of unknown terms:

- *cave diving → jalovo potapljanje* /jalovo means barren or futile/, *karst processes → krasni procesi* /krasni means splendid/
- *non-paleokarstic rocks → nepaleokarstične kamnine, non-karst areas → nekarska območja, glaciation → glacijacija, aerially exposed → ajerno izpostavljeni* /nepaleokarstični, nekarska, glacijacija, ajerno are all newly coined words in Slovene)

For the Slovene-English language pair, PMBT makes similar types of errors as described before, but fewer ambiguity-related errors:

- untranslated terms (*nepaleokraške kamnine → nepaleokraške rocks, kompetitorskih vrst → kompetitorskih species, pobočja vadijev → vadijev slopes*)
- wrong or non-terminological translation (*brezstropa jama → roofless cave* /instead of denuded cave/, *jamski rov → underground tunnel* /instead of cave passage/, *udornica → hollow, precipice, collapsed, sinkhole* /instead of collapse doline/)
- some confusion with geographical names (*reka Reka → river River, Kras → Karst* /instead of Kras when it refers to the Kras plateau/), although great consistency in the translations of *Divaški kras → Divača karst* or *Škocjanske jame → Škocjan Caves.*

NMT from Slovene into English has other types of problems:

- "strange" translations, possibly due to wrong decomposition of the source term (*vrtač → crop*

*rotation* /instead of sinkhole/) or simply inexplicable (*zakraselost → naivety, zakrasele planote → plumed plateaus* /instead of karstification and karstified plateaus/, *melioracija → reclamation*)

- great inconsistencies for the term *udornica* (*udornica → collapse, udder, cliff, collision, burrow, groove* /instead of collapse doline/)
- unsuccessful attempts of generating the correct form of proper names (*Senožeški potok → Senožeški brook, Divaški kras → Divaški karst, Divačski karst; Orehovški kras → Orehovsk karst, Orehovska karst, Orehovsky karst*).

It would appear that lexical choice and disambiguation are still areas where NMT systems have significant room for improvement, despite the fact that NMT translations often indeed appear more fluent or natural than PBMT.

## 4. Discussion and conclusions

It is common wisdom that if we want an MT system to be good at tackling terminology and translating specialised texts, we should train or customize it for the domain of choice. But in many professional translation settings such customization is not easily integrated into the daily workflow, and many freelance translators work in multiple domains. There have been interesting attempts to facilitate such customization and help users "inject" bilingual terminologies into an existing MT system used in a computer-assisted translation (CAT) environment (Arčan et al. 2014). Still, in many cases the "general purpose" MT system is used to translate specialised content without customization.

According to itself, GT serves over 500 million users monthly and translates over 140 billion words per day, which is more than the entire language industry produces in a year[5]. Given these volumes it becomes clear that a considerable portion of this input must be specialised. NMT has clearly improved the fluency of translated output and will likely continue to amaze with methods for the handling of unknown words, it seems however that the accuracy and consistency of terminology still leave room for improvement.

Our evaluation of GT's phrase-based and neural models for the English-Slovene language pair in both translation directions was primarily aimed at testing whether NMT performs better on domain-specific texts, whereby a focused automatic and human evaluation was performed for the accuracy of term translations. A general evaluation with metrics indicates that NMT indeed produces better quality translations in both directions, however for terminology such an improvement was observed only for the English-Slovene translations and not vice versa.

## 5. Bibliographical References


Arčan, M., Turchi, M., Tonelli, S., & Buitelaar, P. (2014). Enhancing statistical machine translation with bilingual terminology in a CAT environment. In *Proceedings of the 11th Biennial Conference of the Association for Machine Translation in the Americas (AMTA 2014)* (pp. 54-68).

Bahdanau, D., Cho, K. and Y. Bengio (2015). Neural machine translation by jointly learning to align and translate. In Proc. of ICLR, San Diego, US-CA.

Doddington, G. (2002). Automatic evaluation of machine translation quality using N-gram CoOccurrence statistics, in Proc. of Second International Conference on Human Language Technology (HLT), San Diego, March 2002, pp. 138-145.

Klubička, F., Toral, A., & Sánchez-Cartagena, V. M. (2017). Fine-grained human evaluation of neural versus phrase-based machine translation. The Prague Bulletin of Mathematical Linguistics, 108(1), 121-132.

Lommel, A., Uszkoreit, H., & Burchardt, A. (2014). Multidimensional Quality Metrics (MQM). *Tradumàtica*, (12), 0455-463.

Machacek, M., and Bojar, O. (2014, June). Results of the WMT14 Metrics Shared Task. In *WMT@ ACL* (pp. 293-301).

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02, pages 311–318.

Peterson, K. and M. Przybocki, NIST 2010 Metrics for Machine Translation Evaluation (MetricsMaTr10) Official Release of Results, http://www.itl.nist.gov/iad/mig/tests/metricsmatr/2010/results

Toral, A., & Sánchez-Cartagena, V. M. (2017). A Multifaceted Evaluation of Neural versus Phrase-Based Machine Translation for 9 Language Directions. arXiv preprint arXiv:1701.02901.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., ... and Klingner, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144.


[5] https://translate.google.com/intl/en/about/