A Sanskrit grammar-based approach to identify and address gaps in Google Translate's Sanskrit-English zero-shot NMT

Amit Rao[1], Kanchi Gopinath
Plaksha University, Mohali, Punjab 140306, India
amitrao.human@gmail.com, gopinath.kanchi@plaksha.edu.in

Abstract

In this work, we test the working of Google Translate's recently introduced Sanskrit-English translation system using a relatively small set of probe test cases designed to focus on those areas that we expect, based on a knowledge of Sanskrit and English grammar, to pose a challenge for translation between Sanskrit and English. We summarize the findings that point to significant gaps in the current Zero-Shot Neural Multilingual Translation (Zero-Shot NMT) approach to Sanskrit-English translation. We then suggest an approach based on Sanskrit grammar to create a differential parallel corpus as a corrective training data to address such gaps. This approach should also generalize to other pairs of languages that have low availability of learning resources, but a good grammar theory.

1 Introduction and motivation

Translation between Sanskrit and English presents significant challenges, even for expert human translation, due to the unique features of Sanskrit and the large linguistic gap between Sanskrit and English. Also, Sanskrit has a unique role, specially in the Indian subcontinent as it was formerly a common language of communication across all fields. It was displaced by colonization, but is seeing a resurgence of late, specially for accessing Indian Knowledge Systems in the original. This has important implications on the expectations from an automatic translation, and the implications of erroneous translations.

In addition to the above factors, automatic machine translation between Sanskrit and English presents the additional challenge of Sanskrit having relatively low availability of high-quality training resources such as tagged corpora.

In May 2022, it was announced (Google, 2022) that 24 new languages including Sanskrit were being added to Google Translate, using the new Zero-Shot Machine Translation adaptation (Johnson et al, 2017) of Google's Neural Multilingual Translation (Wu et al, 2016).

In essence, Zero-Shot NMT leverages deep learning in a single model trained on multiple language pairs, to translate even between directions and language pairs it has not been explicitly trained on. Google's Zero-shot NMT is a variation of Zero-Resource Machine Translation (Firat et al, 2016), which requires an additional fine-tuning step using "pseudo-parallel" data of the new language pair. The need for this step is avoided in the design of zero-shot NMT.

It is interesting that Google has used the Zero-Shot NMT approach for Sanskrit-English. This is a data-driven approach to MT and NLP that avoids the need for explicit encoding of knowledge. The other possible approach to MT and NLP is grammar-based or model-driven, which needs explicit encoding of knowledge. The linguistic theory and grammar behind Sanskrit is very stable and sound. Using this theory base, efforts have been made to create grammar-based Sanskrit NLP systems, for example (Kulkarni, 2021).

The data-driven approach is normally attractive when training resource availability is good, and a good grammar model is absent or too complex. The

[1] Work carried out while the author was at Plaksha.

grammar-based approach is normally attractive when there is a good grammar model, and the training resource availability is poor. With Sanskrit, there is a relatively good grammar model and relatively poor training resource availability. And yet, Google has chosen the data-driven approach for Sanskrit-English translation through its choice of zero-shot NMT, in order to maintain a uniform approach across all languages.

It is pertinent, therefore, to test the effectiveness of Google's Sanskrit-English translation system in actual use. We now describe the considerations we used to design this test.

2 Test Design Considerations for Sanskrit-English Google Translate

We now describe the considerations that came up when designing a test for Sanskrit-English Google translate, and how we dealt with them:

a. **Directionality** – Should we test Sanskrit to English, or English to Sanskrit or both? For the Sanskrit-English pair, we expect three main use cases:
   A. **Sanskrit-English for Sanskrit access** - People fluent in English and interested in Sanskrit literature trying to translate a traditional Vedic or classical Sanskrit mantra, shloka, poem or text from Sanskrit to English.
   B. **English-Sanskrit for Sanskrit learning or communication** - People fluent in English, interested in learning Sanskrit (either for conversation or to access the literature) and trying to translate English to Sanskrit.
   C. **Sanskrit-English for communication or learning** - People fluent in Sanskrit but not English, translating their original Sanskrit text into English for communicating with others or to learn English.

On account of the nature of the language pair and their current status, of the above three, we expect case A to overwhelmingly dominate, case B to be next, and case C to be relatively insignificant. Thus, the Sanskrit-English direction is the highest priority, and also has higher demands on accuracy, since it is more likely to be dealing with classical texts whose incorrect interpretation could have undesirable cultural consequences. We therefore put more emphasis on testing Sanskrit-English than English-Sanskrit. Most of our discussion will also be about Sanskrit to English translation, unless otherwise mentioned or obvious from the context.

b. **Purpose** - Our purpose is to check whether Google's zero-shot NMT automatic translation is robust enough for the Sanskrit-English language pair, rather than an exhaustive performance analysis of the translation. Therefore, rather than a test suite aimed at complete coverage of the language pair, we will create a probe test suite of a few carefully hand-crafted cases, leveraging grammar and language theory, mainly focused on areas where we expect challenges for translation.

c. **Automation** - Since the test set was small and for one-off use, and to be hand-crafted leveraging human expertise, it was simpler to do it iteratively and manually for now rather than invest effort in automating it.

d. **Sourcing** - We did not find a readily available translation test suite for Sanskrit-English focused on testing the robustness of zero-shot NMT, so we created our own.

3 Test process and results

Based on the considerations discussed above, a small "probe" test suite of approximately 120 test cases was hand-crafted and applied iteratively.

Sanskrit to English translation	**English to Sanskrit translation**
98 test cases	31 test cases

Table 1: Number of test cases

The test cases are not all independent, many are part of a group of inputs iteratively designed to test different variants of a specific area being tested. For example, correct translation of single/dual/plural number involves inputs containing various combinations of these. It is difficult to enumerate the groups, since there are sometimes overlaps where a single test case is logically part of multiple groups. Hence, the above table lists the number of individual test cases and not groups.

Each test case was manually translated by one of the authors, who is fluent in both Sanskrit and English, to create the expected reference output. The test case was then input to Google Translate and the output recorded against it. The output was manually evaluated using a 3-way rating system defined by us as follows.

Rating	Meaning
Ok	**Correct** - Output either matches reference output exactly, or is close enough and there is no change in meaning.
?	**Dubious -** Output is acceptable, but not ideal, and/or translation is not consistent across the group.
*	**Incorrect** - Output is totally unacceptable, as it conveys a totally unintended meaning.

Table 2: Rating system

All the test cases, the expected and actual outputs and ratings are available in the sheet attached as Appendix B – All Test Cases. The ratings are annotated with explanations in a Remarks column where needed.

The results of the test are summarized below.

Sanskrit to English translation		
Rating	**Count**	**Percent**
Ok	37	37.76
?	34	34.69
*	27	27.55
Total:	**98**	
English to Sanskrit translation		
Rating	**Count**	**Percent**
Ok	10	32.26
?	10	23.36
*	11	35.48
Total:	**31**	

Table 3: Test results summary

As seen from the above table, if we take the stricter criterion of only Ok-rated outputs as correct, the accuracy of Google Translate for our probe test is 37.76 per cent for Sanskrit to English, and 32.26 per cent for English to Sanskrit. If we take the more relaxed criterion of only *-rated outputs as incorrect, the accuracy is 72.45 per cent for Sanskrit to English, and 64.52 per cent for English to Sanskrit. Since this is only a probe test, and not an exhaustive coverage test, we cannot claim these as the actual accuracy figures, but the test probe does reveal that there are significant gaps in the performance of Google Translate for both the directions that need to be fixed before the translation can be considered robust. The detailed remarks about each output can be found in Appendix B – All Test Cases. In the following section, we summarize the key observations and their implications.

It must be noted here that Google Translate by design is a learning product and is therefore being continuously updated. The test results are valid as of the time they were conducted, namely, in the third week of December 2022.

4 Key observations and implications

Looking at the test case outputs, we find that given the multiple inherent challenges of Sanskrit-English translation, the system performs surprisingly well for a zero-shot NMT that has possibly not been trained on a single input specific to the Sanskrit-English pair. Of the specific challenge areas tested by the probe test, it does cover quite a wide spectrum of phenomena satisfactorily, in at least a few cases, including *sandhi, samāsa, taddhita,* dual number, three grammatical genders, and the phenomenon of *sati-saptamī* (absolutive locative clause).

Having said that, the output is often inconsistent across variations of a language feature, and dubious or incorrect in several cases.

For instance:

- It fails to disambiguate the word भवति[*bhavati*], based on the context, as the sambodhana (vocative) form of भवती[*bhavatī* - “lady”] rather than the third-person present tense form of the धातु[dhatu](verbal root) भू[*bhū*](to be/become).
- It fails to disambiguate the word नेत्रे[*netre*], based on the context, as the accusative case dual number form of the neuter gender noun नेत्र[netra](eye) rather than its locative case singular number form.

- It incorrectly uses Hindi/Urdu words such as कुर्सी, मेज़, and मौसम, which are totally absent in Sanskrit, rather than the corresponding Sanskrit words, to translate English words such as "chair", "table" and "weather" respectively.
- It fails to split the sandhi correctly in अहं ब्रह्मास्मि[*ahaṃ brahmāsmi*] based on the context, which causes it to interpret it as ब्रह्मा[*brahmā*](Brahma the creator) rather than ब्रह्म[*brahma*](Brahman, the Supreme Self), though it does recognize the distinction correctly in non-sandhi cases.

The above examples have been described and discussed in Appendix A – Example Test Cases.

Based on all the test cases listed in Appendix B, we summarize our analysis of the results of the probe test as follows:

A. Google Translate's zero-shot NMT for the Sanskrit-English language pair covers a fairly broad spectrum of translation phenomena. However, our probe test has revealed several significant gap areas that need to be addressed before the system can be considered robust and reliable for general use.

B. The gaps we have identified are largely a consequence of two system factors falling out of the zero-shot NMT data-driven approach –

   a. Not leveraging linguistic knowledge explicitly, due to the design decision of NMT (Wu et al, 2016)

   b. Not training with language-pair specific data, due to zero-shot usage to deal with low availability of resources (Johnson et al, 2017)

   For example, the training data may have included only translation data for Sanskrit-Hindi/Urdu and Hindi/Urdu-English. As a consequence of these, the system makes the following main types of errors: (a) Inconsistent translation across variants of the same phenomenon (b) The system sometimes erroneously translates into Hindi/Urdu words that do not exist in Sanskrit. (c) Unrecognized Sanskrit words are translated to the nearest similar sounding word seen in the training, which leads to errors.

C. These systemic gaps can be addressed by leveraging grammar and language theory. In particular, Sanskrit has an extremely well-developed grammar and language model that allows for precise and accurate representation of the meaning of a sentence.

5 Characterizing the gaps

The current gaps in Google Translate's English-Sanskrit translation, summarized in the previous section, can be classified into two categories:

1. **Learning gaps** - These are gaps that can be addressed by better training of the zero-shot NMT, by feeding more training data, or tuning the learning parameters. For example, if a specific English idiom is not currently learnt as an idiom, it could be learnt by feeding in examples of its usage in the English-Hindi/Urdu translation corpus. We can expect such gaps to gradually reduce over time as the system is fed with more training data, without any change to the basic zero-shot NMT approach. However, with the approach being suggested in this work, this gap reduction could be speeded up.
2. **Systemic gaps** due to pure zero-shot NMT - These are gaps that arise due to not feeding translation data specific to the target language pair (in our case English-Sanskrit) in training the system, but only leveraging translation data of other language pairs that between them cover the target language pair, for example, in our case, English-Hindi/Urdu and Sanskrit-Hindi/Urdu. Such gaps are inherent to the pure zero-shot NMT approach and will not reduce over time. Addressing such gaps needs a different approach that we shall touch upon shortly.

Let us try to formally characterize these two types of gaps in order to understand them better. In order to do that, we must first characterize different types of machine translation systems and see where zero-shot NMT fits in.

Essentially, a deep-learning based machine translation system is a transformer that takes a text s in the source language S and transforms it into a text t in the target language T. In order to do this, it uses a pre-trained language model.

Let **D(L1, L2)** represent training data consisting of parallel translations from the language L1 to the language L2.

Google Translate's Neural Multilingual Translation (NMT) approach (Wu et al, 2016) uses a single Large Language Model (LLM) that is trained on all the languages available. This allows learning to be leveraged across different data sets and languages. Thus, for example, the learning from D(L1, L2), D(L2, L3) and D(L4, L2) is merged into a single model. Then, in translating (L1, L2), the learning from not just D(L1, L2), but also D(L2, L3) and D(L4, L2) gets leveraged, leading to more robust output than from just D(L1, L2) alone. Moreover, this allows the system to translate even between language pairs even though that specific pair was not part of the learning data, say, due to low availability of training data for that pair. In this case, for example, the system could give a translation for (L1, L4), though this pair was not part of the training data. This would obviously not be as robust as having D(L1, L4) in the training mix, but may be better than giving no translation at all. This is what Google means by zero-shot NMT (Johnson et al, 2017), and that is what is reportedly used in English-Sanskrit translation. The implication is that the system is currently not trained on D(English, Sanskrit) data at all. It leverages, for example, D(English, Hindi/Urdu) and D(Sanskrit, Hindi/Urdu), and possibly data in other language pairs, to attempt (English, Sanskrit) translation.

Let **F(S,T)** be the set of features that would have been learnt by the NMT if it had been trained with the ideal training data set D(S,T) to correctly perform an arbitrary (S,T) translation request. Now, in zero-shot NMT, there is no D(S,T). Instead, the NMT is fed an *n*-member set D_n of D(L_i, L_j) where $1 <= i <= n$, $1 <= j <= n$, and (L_i, L_j) is not in D_n. The assumption here is that F(S,T) will be compositionally learnt by the NMT via some combination of D(L_i, L_j) training inputs.

We can now characterize the two categories of gaps we mentioned above in these terms.

A **learning gap** is one where F(S,T) is not currently achieved, but can be learnt by either adding more data to Dn, or by optimizing the parameters of the NMT, or both. For example, let us say a specific form of a verb in Sanskrit is not being correctly translated into English. This could be addressed by adding data that contains that form in the D(Sanskrit, Hindi/Urdu) set.

A **systemic gap** is one where no combination of D(L_i, L_j) comprising D_n for any value of *n* can cause learning of F(S,T), because there exists a set of features **Fo(S,T)** that are not compositional, but can only be learnt by training on D(S,T).

Let us look at an example each of both these gaps.

First, an example of a learning gap. The word मां[*māṃ*] occurs both in Sanskrit and Hindi/Urdu. In Sanskrit, it is the sandhi form of माम्[*mām*] and means "me". In Hindi/Urdu, it is the simplified form of माँ[mām̐] and means "mother". Currently, Google Translate sometimes confuses these two cases, as seen from some of the test outputs. This distinction can be trained into the system by having more examples distinguishing the two cases in the D(Sanskrit, Hindi/Urdu) and D(Hindi/Urdu, English) training data. Hence, we can call this a learning gap.

Now, an example of systemic gap. The unambiguous Sanskrit sentence "अश्वे उपविशन् पिता पुत्रं पश्यति"[*aśve upaviśan pitā putraṃ paśyati*] is ideally translated into English as "The father seated on a horse sees his son". Similarly the unambiguous Sanskrit sentence "पिता अश्वे उपविशन्तं पुत्रं पश्यति"[*pitā aśve upaviśantaṃ putraṃ paśyati*] is ideally translated into English as "The father sees his son **who is** seated on a horse". In Sanskrit, the distinction between the two is clear and marked by inflection on the appropriate noun. In English, the distinction is clear when marked with a relative clause marked by "who is". However, a possible Hindi translation of both these would be "पिता अपने बेटे को घोड़े पर बैठे हुए देखता है"[*pitā apane beṭe ko ghoḍe para baiṭhe hue dekhatā hai*]. This literally translates to "The father sees his son seated on a horse", which is ambiguous due to the prepositional phrase attachment ambiguity, and can convey both the meanings. The issue here is that the unambiguous case markings of Sanskrit on the phrase "seated on the horse" get mapped in both the Hindi translations to a single oblique case marking (बैठे हुए[*baiṭhe hue*]), which has the effect of saying "seated on a horse" without adding the relative clause marker "who is". This feature mismatch (divergence pattern) between the languages (namely inflection in Sanskrit vs oblique case in Hindi vs relative clause in English) causes a non-compositionality in translation. Therefore there is an information loss in zero-shot NMT with transfer learning involving only D(Sanskrit, Hindi/Urdu) and D(Hindi/Urdu, English). This gap cannot be addressed by adding any amount of D(Sanskrit, Hindi/Urdu) and D(Hindi/Urdu, English) data or tweaking the parameters of the zero-shot NMT. It can only be addressed by adding examples of both Sanskrit sentences translated correctly directly to English, that is, by adding some D(Sanskrit, English) data. Thus, we can call this a systemic gap. [Note: This example is slightly simplified for ease of understanding. The actual Sanskrit-Hindi and Sanskrit-English translations by Google Translate are marginally different, but close enough for the example and the argument to hold. The actual details are discussed in Appendix C – "The Systemic Gap Example"].

Systemic gaps are an outcome of "language divergence" in translation, which was formally described in (Dorr, 1994). A partial set of language divergence patterns between English and Sanskrit was described in the context of a prototype rule-based machine translation for English to Sanskrit in (Mishra and Mishra, 2009).

The conclusion we can draw is that due to the presence of systemic gaps, the purely data-driven approach of zero-shot NMT, without encoding grammar knowledge, and without training on the specific language pair, does not work well enough for the Sanskrit-English language pair.

The zero-resource approach of (Firat et al, 2016) using "pseudo-parallel" Sanskrit-English data will not work in the presence of systemic gaps. Google's zero-shot NMT paper (Johnson et al, 2017) discusses (in section 4.6, "Zero-Shot Translation" and section 4.7, "Effect of Direct Parallel Data") possible approaches to go beyond zero-shot NMT by adding real parallel data in the missing language pair and direction (e.g. Sanskrit-English in our case). If a lot of high-quality real-life parallel data is available, this is found to be ideal. However, we know that high-quality Sanskrit-English real-life data is not readily available. For example, Samanantar, a parallel corpora collection between English and 11 Indic languages, does not include Sanskrit (Ramesh et al, 2022). A more recent effort, IndicTrans2, includes parallel corpora with English and all 22 scheduled Indian languages including Sanskrit (AI4Bharat et al, 2023). This is a good start, however, the size of the corpus is only of the order of 2.5k real-life sentences, which may be insufficient for good quality learning. Also, the direction is English-Sanskrit, and reversing it for Sanskrit-English training would not be ideal.

To address this need for direct parallel Sanskrit-English data to plug the gaps in zero-shot NMT, and the absence of such large high-quality corpora, we suggest the following approach – construct a "differential corpus" as a corrective data suite of Sanskrit-English language-pair-specific training data, by leveraging such linguistic knowledge. After all, we leveraged linguistic knowledge to create an effective probe test without requiring an exhaustive test suite. In the same way, we propose that this linguistic knowledge could be leveraged to create the right training data to address these areas and plug the gaps to create a more robust zero-shot NMT system.

To our knowledge, there is no readily available framework that can be directly leveraged to create such a differential corpus. The work on language divergence in general (Dorr, 1994) and English-Sanskrit (Mishra and Mishra, 2009) in particular cited earlier is a good directional starting point, but we will need to focus it on the Sanskrit-English direction and cover a more comprehensive set of features specific to the systemic gap issue than addressed in those works. There has been work done in identifying the formal structure of Sanskrit text (Huet, 2009) that could provide a set of features for us to take into account, but they may need to be filtered to keep only the ones relevant from the viewpoint of divergence. There is also a translation of the exercises from Apte's classical text on Sanskrit syntax (Apte, 1885) into Sanskrit (Gillon, 1996). This again may need to be filtered for divergence specific cases. A generic grammar framework like The Grammatical Framework (GF, 1998) is potentially interesting for its ability to deal with multilingual grammars, but it currently has support only for English and Hindi, not Sanskrit. Sanskrit-specific tools and toolkits such as Inria's Sanskrit Heritage Site (Huet, 1994), the University of Hyderabad Department of Sanskrit's Samsaadhani (Kulkarni, 2002) and Ashtadhyayi.com (Bodas, 2015) are targeted at understanding Sanskrit rather than translating it to English, so they may be useful as reference tools for the human experts generating the differential corpus.

Large language models (LLMs) like ChatGPT (Yiheng Liu et al, 2023) get their capabilities through pre-training, an expensive and long drawn-out process. Reinforcement Learning from Human Feedback (RLHF) (Ziegler et al, 2019) can be used to affect their behaviour (or to "finetune" it, such as eliciting right responses or suppressing objectionable output), which is not as expensive, and hence can be undertaken multiple times. RLHF can potentially be used with the differential corpus suggested, but this needs further study.

In-context learning (Sang et al, 2022) is seen as a major shift in transfer learning in the context of LLMs, with intimations of an "emergent" behaviour. In contrast to the classic pretraining-then-finetuning procedure for downstream prediction tasks in LLMs, there is only a need to provide a few "in-context" examples, without affecting existing model parameters. The differential corpus suggested might well function as in-context examples that can be taken up as future work.

The proposed differential corpus would consist of D(Sanskrit, English) and D(English, Sanskrit) data that would be designed, based on knowledge of Sanskrit and English grammar theory, to focus on addressing the systemic gap, that is, exercising the language features Fo(Sanskrit, English) and Fo(English, Sanskrit) that are not compositionally learnable from D(S,T) data sets not containing the above two language pairs.

In addition, since we are going to cover all combinations of a given feature, in the process, it may incidentally cover some learning gap data as well, because, intuitively, if a feature can be compositionally learnt from $D(S, L_1)$ and $D(L_2,T)$, then it can also be directly learnt from D(S,T). Thus, the suggested approach would also speed up reduction of the learning gap.

6 Outline of grammar-based approach for identifying a "differential corpus" as corrective training data

The key idea behind the proposed grammar-based approach is to leverage the rich linguistic model of Sanskrit from the traditional Indian *śāstra*s including vyākaraṇa (the aṣṭādhyāyī of Pāṇini and its related works), the vākyapadīyam of Bhartṛhari, mīmāṃsā, nyāya and vaiśeṣika, mapping approximately to linguistics, grammar, discourse analysis, logic and ontology respectively, to create a "differential corpus" of translation test cases that can be used as training data to fill the current gaps for Sanskrit-English zero-shot NMT.

The proposed approach is summarized as follows:

A. Identify the prominent divergence areas of the Sanskrit-English language pair, that is, the set of language features that are present in Sanskrit and either absent or rudimentary in English, Fo(Sanskrit, English) or vice versa, Fo(English, Sanskrit). For reasons stated earlier, we focus here on the first case, Fo(Sanskrit, English).

B. For each feature, iteratively create a group of test cases to test the translation of that feature. A group consists of a set of individual test cases. Collectively, the group should cover the range of variations of that feature. For example, if the divergence area is – presence of three grammatical numbers (singular/dual/plural) in Sanskrit, versus only two grammatical numbers (singular/plural) in English - the feature is "grammatical number", and we have to create as test input a group of sentences containing all combinations of singular, dual and plural nouns.

Of particular interest are cases of ambiguity, where two or more features map to the same form (e.g. a tiṅanta and subanta, or a kṛdanta and subanta, map to the same form as seen in the भवति example). The test inputs should check whether the translation deals with the ambiguity and provides the correct translation.

Such a differential corrective parallel corpus can be fed to the existing zero-shot NMT in addition to the training data it has already seen, without the need for any significant modification to the architecture of the system.

We now identify the key linguistic features of Sanskrit that are part of the proposed approach as outlined above, and highlight the potential challenge areas to be tested in each.

A. Lexical features

1. Sandhi - correct identification of all स्वर[svara](vowel), व्यञ्जन[vyañjana](consonant) and विसर्ग[visarga](aspirant) sandhis. Particularly where sandhi leads to ambiguous forms. For example: ब्रह्मास्मि[brahmāsmi] can be ब्रह्म

अस्मि[brahma asmi] or ब्रह्मा अस्मि[brahmā asmi].

2. **Special signs** - such as the अवग्रह-चिह्न[*avagraha-cihna*]ऽ.
For example: अनुगृहितोऽस्मि[*anugṛhito'smi*] / अनुगृहितोस्मि[*anugṛhitosmi*] / अनुगृहितः अस्मि[*anugṛhitaḥ asmi*] are equivalent.

B. Morpho-syntactic features

3. सुबन्त[subanta](noun) forms - correct handling of ambiguous forms such as ते.

4. Basic तिङन्त[tiṅanta](verb) forms - correct handling of same धातु[dhātu] in multiple गणs[*gaṇa*s](groups)having same forms with different meanings, or having same form as subantas (e.g. भवति[bhavati]).

5. Derived tiṅanta(verb) forms - correct handling of verbs derived from णिच्[ṇic], सन्[san] and similar प्रत्ययs[pratyayas] (suffixes).

6. Compound sentences - correct handling of यद्/तद्[yad/tad] and similar conjoint sentences.

7. Complex sentences - correct handling of clauses involving कृदन्त[kṛdanta] (participials).

C. Semantic features

8. Word order and topicality - word order does not change the gross meaning in Sanskrit, but may alter the focus and topicality. Also, in some cases the order does matter, for example, placement of अपि[api] at the beginning vs middle.

9. तद्धित[taddhita] (noun-noun morphology) - for example, अण्[aṇ] patronymic pratyaya

10. समास[samāsa] (compound nouns) - correct translation of all the main samāsa types - तत्पुरुष[tatpuruṣa](including कर्मधारय[karmadhāraya], द्विगु[dvigu], उपपद[upapada], and नञ्[nañ]), बहुव्रीहि[bahuvrīhi], अव्ययीभाव[avyayibhava] and द्वन्द्व[dvandva]. For example – the same samāsa (e.g. पीताम्बर) can be interpreted as तत्पुरुष[tatpuruṣa] or बहुव्रीहि[bahuvrīhi] depending on the context.

A parallel corpus based on the above feature set could be created by leveraging the related work discussed earlier, as well as taking example sentences given for different grammar features of Sanskrit from a good Sanskrit grammar book, for example, (Rao, 2022), extrapolating them for complete coverage of all variations, and providing English translations. In some cases, the sentences may have to be hand-crafted as we have done here. Since we are looking at only a differential corpus, we estimate the number of test groups to be of the order of approximately a thousand in number, which is feasible to do manually in a reasonable time frame.

We believe a "differential" corrective translation data suite based on this model will allow most of the gaps in Google Translate's zero-shot NMT for Sanskrit-English to be addressed, leading to a more robust and usable system.

7 Contributions and scope for future work

This work has made the following contributions:

1. Through a small hand-crafted probe-test suite, we have shown that though Google Translate's recently introduced Sanskrit-English service based on zero-shot NMT covers a broad spectrum of cases adequately, there are still significant gaps in translation performance.
2. We have identified that the gaps are either learning gaps due to inadequate training data or need for parameter tuning, or systemic gaps due to the nature of zero-shot NMT and the divergence between the languages, and both these can be addressed by leveraging Sanskrit linguistic knowledge available in traditional works such as Bhartrhari's Vakyapadiyam (Sharma, 2016), and the *vyākaraṇa, mīmāṃsā* and *nyāya-vaiśeṣika* traditions that it references.
3. We have proposed an approach for creating corrective translation data for Sanskrit-English translation to address the systemic

gaps identified. We believe this idea is generalizable to other language pairs where there is a divergence in the language pair and a rich linguistic knowledge base exists.

Scope for future work includes:

1. Extending the approach to include English to Sanskrit direction considerations.

2. Creating an actual differential corrective translation test suite based on the approach.

3. Applying the differential corrective suite to Google Translate and measuring its impact.

Limitations

This work is subject to the following known limitation:

The solution proposed is currently indicative and directional, based on a theoretical understanding of how Google's zero-shot NMT works, and how the demonstrated gaps may have arisen, based on published literature on the system, and the authors' experience with translation and linguistics. The authors have not had access to the source code of the system, or any involvement with the actual training of the system. The proposed solution needs to be detailed out and practically implemented, ideally in collaboration with Google.

Ethics Statement

To the best of the authors' knowledge and belief, this work is fully compliant with the ACL Ethics Policy. We have identified significant gaps in the current working of zero-shot NMT for the Sanskrit-English pair, and have made suggestions for addressing these gaps. We believe these suggestions, if successfully implemented, will lead to a more robust and accurate system, thus improving the state of the art, which will benefit all stake-holders.

References

Google Blog. 2022. Google Translate learns 24 new languages. *Google IO announcement on Twitter*. https://twitter.com/Google/status/1524438946293374986.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi et al. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *Computing Research Repository*, arXiv1609:08144. Version 2.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat et al. 2017. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *Computing Research Repository*, arXiv1611:04558. Version 2.

Mike Schuster, Melvin Johnson, and Nikhil Thorat. 2016. Zero-Shot Translation with Google's Multilingual Neural Machine Translation System. *Google Research Blog post.* https://ai.googleblog.com/2016/11/zero-shot-translation-with-googles.html.

Amba Kulkarni. 2021. Sanskrit Parsing Following Indian Theories of Verbal Cognition. In *ACM Transactions on Asian and Low-Resource Language Information Processing,* Volume 20, Issue 2, March 2021. Article No. 23, pages 1-38. https://dl.acm.org/doi/10.1145/3418061.

Raghunath Sharma. 2016 edition. वाक्यपदीयम् – Vakyapadiyam of Bhartrhari with Two Sanskrit Commentaries. Sampurnanand Sanskrit University, Varanasi, India. 2016 edition. ISBN 978-81-7270-020-1.

Tilaka Rao. अष्टाध्यायीप्रवेशः[*aṣṭādhyāyīpraveśaḥ*] vol.1 &2. 2022. Siddhanta Knowledge Foundation, Chennai, India. 2022 edition. ISBN 978-81-9569-142-5 and 978-81-9569-143-2.

Bonnie Dorr. 1994. Machine Translation Divergences: A Formal Description and Proposed Solution. *Computational Linguistics* 20(4):597-633.

Mishra, V. and Mishra, R. B. 2009. Divergence patterns between English and Sanskrit Machine Translation. *INFOCOMP Journal of Computer Science* 8(3): 62–71.

Huet, G. 2009. Formal Structure of Sanskrit Text: Requirements Analysis for a Mechanical Sanskrit Processor. In: Huet, G., Kulkarni, A., Scharf, P. (eds) *Sanskrit Computational Linguistics*. ISCLS ISCLS 2007-2008. Lecture Notes in Computer Science, volume 5402. Springer, Berlin, Heidelberg.

Pawan Goyal and Amba Kulkarni. 2014. Converting Phrase Structures to Dependency Structures in Sanskrit. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers:1834-1843, Dublin, Ireland.*

Orhan Firat, Baskaran Sankaran, Yaser Al-onaizan, Fatos T. Yarman Vural, and Kyunghyun Cho. 2016. Zero-Resource Translation with Multi-Lingual Neural Machine Translation. In *Proceedings of the*

2016 Conference on Empirical Methods in Natural Language Processing, pages 268–277, Austin, Texas. Association for Computational Linguistics.

Vaman Shivaram Apte. 1885. The Student's Guide to Sanskrit Composition. A Treatise on Sanskrit Syntax for Use of Schools and Colleges. Lokasamgraha Press. Pune, India.

Brendon S Gillon. 1996. Word Order in Classical Sanskrit. Indian Linguistics, 57(1-4):1-35.

Gowtham Ramesh, Sumanth Doddapaneni and 16 other authors. 2022. Samanantar: The Largest Publicly Available Parallel Corpora Collection for 11 Indic languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

AI4Bharat and 14 other authors. 2023. IndicTrans2: Towards High-Quality and Accessible Machine Translation Models for all 22 Scheduled Indian Languages. *arXiv preprint arXiv:2305.16307* (2023).

GF. 1998. Grammatical Framework. A programming language for multilingual grammar applications. https://www.grammaticalframework.org/.

Gerard Huet. 1994. The Sanskrit Heritage Site. https://sanskrit.inria.fr/index.fr.html .

Amba Kulkarni. 2002. संसाधनी[*saṃsādhanī*]. A Sanskrit Computational Toolkit. https://sanskrit.uohyd.ac.in/scl/ .

Neelesh Bodas. 2015. ashtadhyayi.com. A comprehensive study resource for the students of Panini grammar. https://ashtadhyayi.com/ .

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2022. An Explanation of In-context Learning as Implicit Bayesian Inference. *arXiv:2111.02080 [cs.CL]* ICLR 2022.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, Geoffrey Irving. 2019. Fine-Tuning Language Models from Human Preferences. *arXiv:1909.08593* .

Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, Zihao Wu, Dajiang Zhu, Xiang Li, Ning Qiang, Dingang Shen, Tianming Liu, Bao Ge. 2023. Summary of ChatGPT/GPT-4 Research and Perspective Towards the Future of Large Language Models. *arXiv:2304.01852 [cs.CL]* .

Appendix A – Example test cases

Example 1 (Ok):

Consider the following example of Sanskrit to English translation by the system. Though it was not a part of the probe test and therefore not in the appendix, it was the sentence that piqued our interest in carrying out this probe test.

Test input	**Reference (expected) output**	**Google Translate output**	**Rating and remark**
भ्रम्(१/४) धात्वोः विषये भ्रमः(४) मा भूत् इति व्याख्यानप्रपञ्चे वयं भ्रमन्तः(१) स्मः। [*bhram(1/4) dhātvoḥ viṣaye bhramaḥ(4) mā bhūt iti vyākhyānaprapañce vayaṃ bhramantaḥ (1) smaḥ*]	We are wandering(1) in the world of explanation so that there should be no confusion(4) about the verb bhram(1/4).	We are wandering(1) in the world of explanation so that there should be no confusion(4) about the verb bhram(1/4).	Ok – correctly disambiguates the two senses of bhram (wandering and confusion) based on the context.

Table 4: Example 1 (Ok)

This sentence is potentially challenging to translate,because the dhaatu(verbal root) भ्रम् [bhram] in Sanskrit occurs in two *gaṇa-s* (verb groups) (namely, 1 and 4), with different connotations, namely, “to wander” in group 1, and “to be confused” in group 4. Many of their forms are similar. Therefore, the sentence involves wordplay and ambiguity, which is traditionally a challenge for translation. However, Google Translate correctly translates these two senses. We could therefore say that the system has “learnt” the two senses of the verb root. However, with deep learning, such inputs are never explicitly encoded. Moreover, with zero-shot NMT, it is likely that the system was never fed an instance of Sanskrit to English translation of either of these senses. And yet, the system exhibits learning behaviour for this translation pair. This illustrates the power of deep learning and zero-shot NMT that causes learning without structuring and encoding of knowledge.

Along similar lines, the probe test analysis reveals that the zero-shot NMT has “learnt” a number of language phenomena that are potentially challenging for Sanskrit-English, in at least a few cases - *sandhi, samāsa, taddhita,* dual number, 3 grammatical genders, and the phenomenon of *sati-saptamī* (a kind of absolutive locative clause).

On the other hand, the system gives dubious, incorrect or inconsistent output for a number of cases, as seen in the remaining examples.

Example 2 (Error):

The word form भवति[*bhavati*] in Sanskrit is [ambiguous between the vocative form of the noun

Test input	Reference (expected) output	Google Translate output	Rating and remark
भवति भिक्षां देहि। [bhavati bhikṣāṃ dehi]	Madam, give me alms.	Give me the alms you have.	* - Does not recognize भवति[bhavati] as the sambodhana (vocative) of भवती[bhavatī - "lady"] .

Table 5: Example 2 (Error)

भवती[*bhavatī*](lady), and the simple present tense of the verb भू[*bhū*](to be/become). The verbal form is much more common in use than the vocative noun form. A grammar-based analysis would be able to deal with this ambiguity using knowledge-based disambiguation rules; a data-driven system would pick the statistically more common meaning, unless specifically exposed to this instance, which is highly unlikely with zero-shot NMT. This test case is a famous sentence from the famous epic Ramayana. Sentences from the epics are likely to be commonly queried for Google Translate, and getting it wrong is a fairly serious gap.

Example 3 (Dubious):

Test input	Reference (expected) output	Google Translate output	Rating and remark
नेत्रे पश्यतः। [netre paśyataḥ]	(Two) eyes see.	Looking into the eyes.	? - Confuses dual-number neuter-gender form with the *saptamī-vibhakti* (locative case) form.

Table 6: Example 3 (Dubious)

Similarly, the word form नेत्रे [*netre*] is ambiguous between dual number nominative/accusative case, or singular locative case. Here again, in the context of the word पश्यतः [*paśyataḥ*], the overall sentence is unambiguous in the light of Sanskrit grammar, however zero-shot NMT currently fails to get it right. This is not a classical sentence, but a simple Sanskrit sentence that is expected to be correctly translated.

There is another point this example serves to illustrate. In translation from language A to language B, if language A has a feature F that is absent in language B, and we are translating a sentence from A to B that involves the use of this feature F, then by default the canonical translation into B will lead to loss of information of the feature F. In this case, for example, since Sanskrit has the dual number while English does not, the sentence with the dual number would be folded to the plural number in English, thus leading to loss of information. One way to deal with this is to explicitly insert this information in some way in the target language, as we have done by adding "two" in parentheses in the reference expected output for this sentence. Whether to do this or not is a matter of choice, but the choice should be exercised uniformly for consistency. Examining the probe test cases in detail, we find that since zero-shot NMT does not explicitly deal with encoding any language feature such as number, the output is inconsistent, and depends on the training data instances that it has been exposed to.

Example 4 (Inconsistent):

Test input	**Reference (expected) output**	**Google Translate output**	**Rating and remark**
अहं ब्रह्म अस्मि। *[ahaṃ brahma asmi]*	I am the Brahman.	I am the Brahman.	Ok - Understands distinction of ब्रह्मन् (Brahman - neuter gender word) vs ब्रह्मा (Brahma - masculine gender word).
अहं ब्रह्मा अस्मि। *[ahaṃ brahmā asmi]*	I am Brahma.	I am Brahma.	Ok
अहं ब्रह्मास्मि। *[ahaṃ brahmāsmi]*	I am the Brahman.	I am Brahma.	? - Does not handle the ambiguity due to *dīrgha-sandhi* correctly, leading to incorrect output.

Table 7: Example 4 (Inconsistent)

अहं ब्रह्मास्मि [*ahaṃ brahmāsmi*] is an iconic sentence from the Upanishads and is considered a महावाक्य [*mahāvākya*] (great statement) of the Vedic literature, Sanatana Dharma and Hinduism.

It should translate to "I am Brahman" (the Ultimate Reality) and not "I am Brahma" (the four-headed Creator of the world, one of the Trinity of Brahma, Vishnu and Shiva). Google Translate is aware of this distinction, as seen from the first two examples, where the words are separated. However, it fails to recognize and translate it correctly when it is combined as a single word using sandhi. This is problematic and needs to be addressed.

Example 5 (Error):

Test input	Reference (expected) output	Google Translate output	Rating and remark
The chair is made of wood.	आसन्दं काष्ठेन निर्मितम्।	कुर्सी काष्ठेन निर्मिता भवति।	* - Uses Hindi/Urdu word कुर्सी for chair.
The table is made of wood.	उत्पीठिका काष्ठेन निर्मिता।	मेजः काष्ठेन निर्मितः अस्ति।	* - Uses Hindi/Urdu word मेज़ for table.
He did not come to work today as he is feeling a bit under the weather.	सः अद्य कार्यं कर्तुं न आगतः यतः सः किञ्चित् अस्वस्थः अस्ति।	सः अद्य कार्यं कर्तुं न आगतः यतः सः किञ्चित् मौसमस्य अधः अनुभवति।	* - Uses Hindi/Urdu word मौसम for weather. Also, does not understand the idiom "under the weather".

Table 8: Example 5 (Error)

The above three examples are from English to Sanskrit translation. In all three examples, English words have been translated using Hindi/Urdu words derived from Persian/Arabic and which are not Sanskrit words. This indicates that the zero-shot NMT was probably trained on Hindi/Urdu single-language data, and/or English-Hindi/Urdu sentence translation data, and this learning has percolated into English-Sanskrit translation. This is highly problematic for users who are trying to learn Sanskrit, as they will pick up words which are not in Sanskrit and assume them to be Sanskrit words.

Appendix B – All test cases

Part 1 of 2: Sanskrit to English

Test Input	Reference (expected) Output	GoogleTranslate Output	Rating	Remark
मम नाम अमितः।	My name is Amit.	My name is Amit.	Ok	Understands basic sentence with implicit copula "is".
अमितः मम नाम।	Amit is my name.	Amit is my name.	Ok	Seems to handle simple word order variation.
अहं मुम्बईतः।	I am from Mumbai.	I am from Mumbai.	Ok	Understands तः pratyaya used in place of fifth vibhakti.
पादोनसप्तवादनम्।	(It is) a quarter to seven.	It was seven o'clock in the morning.	*	Ignored पादोन for quarter-to. Where does morning come from?
पादोननववादनम्।	(It is) a quarter to nine.	Ninety-nine feet.	*	Confused पादोननव with ninety-nine, ignored वादनम् o'clock.
इदानीं पादोनसप्तवादनम्।	It is now a quarter to seven.	It is seven-foot-seven.	*	Ignored इदानीम् (for now), and confused पाद with feet.
पर्वतो वह्निमान् धूमात्।	The mountain is on fire because it has smoke.	The mountain is on fire from the smoke.	Ok	Understands sandhi.
मूकं करोति वाचालं पङ्गुं लङ्घयते गिरिम्। यत्कृपा तमहं वन्दे परमानन्दमाधवम्॥	He whose grace makes the mute speak and the lame cross mountains, I salute that supreme bliss Madhava.	He makes the speechless dumb and the lame cross the mountain I salute Him whose grace is the supreme bliss, Madhava.	?	Does not handle complex यद्/तद् properly?
या कुन्देन्दुतुषारहारधवला या शुभ्रवस्त्रावृता। या वीणावरदण्डमण्डितकरा या श्वेतपद्मासना॥ या ब्रह्माच्युतशङ्करप्रभुतिभिर्देवैस्सदावन्दिता। सा मा पातु सरस्वती भगवती निःशेषजाड्यापहा॥	She whose whiteness would defeat that of the kunda flower and the moon and of snow, who is adorned with white clothes, whose hands are adorned with the staff of a superior veena, who is seated on a white lotus, who is worshipped by gods beginning with Brahma, Achyuta and Sankara, may that goddess Saraswati, destroyer of all illusion, protect me.	She is white with necklaces of kunda and moon and snow and covered with white clothes That woman with her hands adorned with the staff of a veena and her seat on a white lotus She is always worshiped by the gods like Brahma, Acyuta and Shankara. May that holy Saraswati who destroys all illusion protect me	?	Largely ok, but does not handle complex yad/tad construct, does not get some complex samasas right, and does not understand that this is a single long sentence and needlessly splits it up.

अहं ब्रह्म अस्मि।	I am Brahman.	I am the Brahman.	Ok	Understands distinction of ब्रह्मन् (N - Brahman) vs ब्रह्मा (M - Brahma).
अहं ब्रह्मा अस्मि।	I am Brahma.	I am Brahma.	Ok	
अहं ब्रह्मास्मि।	I am Brahman.	I am Brahma.	?	Does not handle the ambiguity due to deergha sandhi correctly, leading to input that is actually incorrect.
रामकृष्णौ आगच्छतः।	Rama and Krishna are coming.	Ramakrishna and Rama are coming.	*	Does not handle द्वन्द्वसमास correctly?
मां तारय।	Save me.	Save me, mother.	?	Confused मां (माम्) with Hindi word माँ (mother).
मा तारय।	Do not save.	Don't save me.	?	Where did the "me" come from?
सा गजगामिनी।	She has a gait like an elephant.	She is an elephant.	*	Does not understand गजगामिनी as a samasa.
रामः पीताम्बरः।	Rama is one whose dress is yellow.	Rama is dressed in yellow.	Ok	Understands bahuvrihi vs tatpurusha use of पीताम्बर.
सः पीताम्बरः।	He is one whose dress is yellow.	He's in yellow.	?	Why the inconsistency "in yellow" vs "yellow-clad" across genders?
सा पीताम्बरा।	She is one whose dress is yellow.	She's yellow-clad.	?	Why the inconsistency "in yellow" vs "yellow-clad" across genders?
सः पीताम्बरं धारयति।	He is wearing a yellow dress.	He is wearing a yellow robe.	Ok	Understands bahuvrihi vs tatpurusha use of पीताम्बर.
भवति भिक्षां देहि।	Madam, give me alms.	Give me the alms you have.	*	Does not recognize भवति as sambodhana of भवती.
हे देवि, भिक्षां देहि।	Madam, give me alms.	O Goddess, give me alms,	Ok	
देवि भिक्षां देहि।	Madam, give me alms.	O Goddess, give me alms,	Ok	
रामः सीतया आगच्छति।	Rama is coming with Sita.	Rama is coming with Sita.	Ok	Understands all variations of

				upapada tritiya vibhakti with सह.
रामः सीतया सह आगच्छति।	Rama is coming with Sita.	Rama is coming with Sita.	Ok	
रामः सहसीता आगच्छति।	Rama is coming with Sita.	Rama is coming with Sita.	Ok	
रामः ससीता आगच्छति।	Rama is coming with Sita.	Rama is coming with Sita.	Ok	
वागर्थाविव सम्पृक्तौ वागर्थप्रतिपत्तये जगतः पितरौ वन्दे पार्वतीपरमेश्वरौ।	For attaining speech and its meaning, I salute the parents of the world, Parvati and Parameshwara, who are connected as speech and meaning.	I salute the two fathers of the world, the Supreme Lord of Parvati, who are as connected as the meaning of speech, for attaining the meaning of speech.		Does not understand वागर्थौ and पार्वतीपरमेश्वरौ as dvandva samasas. Does not understand पितरौ as an ekasesa samasa.
अनादिनिधनंब्रह्म शब्दतत्त्वंयदक्षरम्। विवर्तते अर्थभावेन प्रक्रिया जगतो यतः॥	The timeless eternal Brahman that is all-pervading, and whose transforms into meaning is the process of the world, is the essence of language.	The eternally immortal Brahman, the essence of sound, is the imperishable. Because the process of the world revolves around the sense of meaning	?	Does not understand the alternate meaning of अक्षरम् as pervasive. Does not properly understand yad/tad complex construct.
कर्मण्येवाधिकरस्ते मा फलेषु कदाचन। मा कर्मफलहेतुर्भूर्मा ते सङ्गोऽस्त्वकर्मणि॥	Your authority is over action alone, not over the fruits thereof, be not motivated by fruits of action, nor be attached to inaction.	You have only right to action, never to fruits. Let not the fruits of action be thy motive, nor let thy attachment be to inaction.	Ok	Understands this Bhagwadgita shloka.
अश्वत्थामो हतोहतः। नरो वा कुञ्जरो वा।	Ashvatthama is killed, whether man or elephant.	Ashvatthama was killed. Whether it is a man or an elephant.	Ok	Understands visarga sandhi.
अश्वत्थामो हतोहतः। नरोऽवा कुञ्जरोऽवा।	Ashvatthama is killed, whether man or elephant.	Ashvatthama was killed. A man or an elephant.	Ok	Understands avasarga, but why the change in translation?
अश्वत्थामो हतोहतः। नरोवा कुञ्जरोवा।	Ashvatthama is killed, whether man or elephant.	Ashvatthama was killed. Narova Kunjarova.	*	Does not split pada boundaries cleanly in absence of whitespace.
अश्वत्थामोहतोहतः। नरोवाकुञ्जरोवा।	Ashvatthama is killed, whether man or elephant.	Ashvatthamohatohata. Narovakunjarova.	*	Does not split pada boundaries cleanly in absence of whitespace.
ते गमिष्यन्ति।	They will go.	They will go.	Ok	
ताः गमिष्यन्ति।	They will go.	They will go away.	?	Where did "away" come from?

सः कुर्यात् सदा मङ्गलम्।	May he always do good.	May he always do good.	Ok	
सः क्रियात् सदा मङ्गलम्।	May he always do good.	He is always auspicious from action.	*	Confuses आशीर्लिङ् लकार form with (incorrect) panchami form.
ममोपात्त दुरितक्षयद्वारा श्रीपरमेश्वरप्रीत्यर्थम्।	For the pleasure of Sri Parameshwara through the destruction of evils attained by me.	For the pleasure of Sri Parameshwara through the destruction of evils attained by me.	Ok	Understands idiomatic usage like प्रीत्यर्थम्.
युगं वर्तते।	The age exists.	The age is present.	Ok	
युगे वर्तेते।	(Two) ages exist.	exists in the age.	?	Dual number not handled correctly and consistently.
युगानि वर्तन्ते।	Ages exist.	There are ages.	Ok	
युगम् अवर्तत।	The age occurred.	The era turned around.	?	Does not understand लङ्लकार (past tense) form अवर्तत of आत्मनेपद अकर्मक dhatu वृत्
युगे अवर्तेताम्।	(Two) ages existed.	Let them turn in the age.	*	
युगानि अवर्तन्त।	Ages existed.	The ages passed.	Ok	
नेत्रे पश्यतः।	(Two) eyes see.	Looking into the eyes.	?	Confuses dual number neutral gender form with saptami vibhakti form.
नेत्राभ्यां पश्यतः।	(They two) see with (two) eyes.	Looking at you with your eyes.	*	Where did "at you" and "your" come from?
श्रेयश्च प्रेयश्च मनुष्यमेतस्तौ सम्परीत्य विविनक्ति धीरः। श्रेयो हि धीरोऽभि प्रेयसो वृणीते प्रेयो मन्दो योगक्षेमाद्वृणीते ॥	The good and the pleasant both approach man. The wise, on examining both, chooses the good. The wise prefers the good over the pleasant, the unwise, compelled by material considerations, prefers the pleasant.	The steadfast man distinguishes between these two, the good and the dear. A sober person seeks the best of his dear ones, and a slow person seeks the safety of mystic yoga.	*	Does not understand meaning of योगक्षेम, does not analyze the shloka correctly.
सः नेत्राभ्यां पश्यति।	He sees through (two) eyes.	He looks through his eyes.	?	Not clear if it recognizes dual number here. Where did "his" come from?
अनुगृहीतोऽस्मि।	I am obliged.	I am gracious.	*	Does not understand standard phrase for "thank you" i.e. "I am obliged".

रामे वनं गते कृष्णः नगरं गतवान्।	When Rama went to the forest, Krishna went to the city.	When Rama went to the forest, Krishna went to the city.	Ok	Understands sati saptami.
रामे वनं गते सति कृष्णः नगरं गतवान्।	When Rama went to the forest, Krishna went to the city.	When Rama went to the forest, Krishna went to the city.	Ok	
रामः भोजनं कृत्वा शालां गच्छति।	Rama eats and goes to school.	Rama eats and goes to the shed.	?	Does not understand different meanings of शाला in context.
रामः अशित्वा शालां गतः।	Rama, having eaten, went to school.	Rama went to the shed without eating.	*	Confuses अशित्वा (having eaten) kridanta form, reverses the meaning.
रामः भोजनम् अशित्वा शालां गतः।	Rama, having eaten food, went to school.	Rama ate the food and went to the shed.	*	Seems to get the right meaning of अशित्वा here, but seems to be order-dependent. Also, misses out the nuance of "having eaten" by using "and".
रामः महामनाः।	Rama is (one who has) a great mind.	Rama is a great mind.	?	Does not understand bahuvrihi consistently. Rama (is one who) has a great mind.
रामः कम्बुग्रीवः।	Rama is (one who has) a neck shaped like a conch shell.	Rama is a conchshell.	*	Does not understand कम्बुग्रीव as bahuvrihi.
युधिष्ठिरः अजातशत्रुः।	Yudhishthira (is one who) has no enemies (born).	Yudhisthira is the unborn enemy.	?	Does not understand नञ् बहुव्रीहि samasa.
दाशरथिः रामः विजयते।	Rama, the son of Dasharatha, is victorious.	Rama, the charioteer of Dasaratha, is victorious.	*	Does not understand दाशरथि as taddhita form, son of Dasharatha. Seems to confuse with सारथी (charioteer).
सागरात् आहिमालयं भारतदेशः।	India (is) from the ocean upto the Himalayas.	From the sea to the Himalayas, India.	?	What about elided copula "is" which is mandatory in English?
आहिमालयं भारतदेशः।	India (is) upto the Himalayas.	The Himalayas are India.	*	Inconsistent recognition of

				avyayibhava samasa with आङ्
चिन्ता मास्तु।	Don't worry.	Don't worry.	Ok	
चिन्तामाऽस्तु।	Don't worry.	Let's worry.	*	Does not handle savarna deergha with avasarga.
रामस्य मनः महत्।	Rama's mind (is) great.	Rama's mind is great.	Ok	
रामस्य महामनः।	Rama's mind (is) great.	Rama's great mind.	?	Elided copula "is" not consistently inferred.
चैत्रवैशाखौ वसन्तऋतुः।	Chaitra and Vaishakha (are) the spring season.	Spring in Chaitra and Vaishakha.	?	Does not handle vidheya viseshana consistently well.
चैत्रवैशाखयोः वसन्तऋतुः।	The spring season (is) in Chaitra and Vaishakha.	Spring is the season of Chaitra and Vaishakha.	?	Confused shashti and saptami identical forms.
तस्य गतवैभवः पुनः न आयास्यति।	His lost glory will never return.	His lost glory will never come back.	Ok	
सः गतवैभवः।	He is one whose glory is gone.	He is a lost glory.	*	Does not understand bahuvrihi correctly.
तस्य प्राप्तविद्या महती।	His acquired knowledge is great.	His acquired knowledge is great.	Ok	
सः सम्प्राप्तविद्यः।	He is one who has properly acquired knowledge.	He is an acquired knowledge.	*	Does not understand bahuvrihi correctly.
कौन्तेयस्य अर्जुनस्य सारथिः श्रीकृष्णः।	Sri Krishna is the charioteer of Arjuna, the son of Kunti.	Sri Krishna is the charioteer of Kaunteya and Arjuna.	*	Does not handle taddhita and viseshana correctly.
गाङ्गेयः भीष्मः कौरवाणां सेनापतिः।	Bhishma, the song of Ganga (is) the commander of the army of the Kauravas.	Ganges, Bhishma, the commander of the army of the Kauravas.	*	Does not handle taddhita and viseshana correctly.
न जातु सः गृहं गच्छति।	He never goes home.	He never goes home.	Ok	Understands use of idioms like जातु.
जातु सः गृहं गच्छति।	Sometimes he goes home.	Jata he goes home.	*	But understanding is not consistent across usages.
प्राणवायवः पञ्चधा।	The prana-vayus are fivefold.	The prana-vayu is fivefold.	?	Understands धा pratyaya following N to mean Nfold, but not plural of वायवः?

रसाः षोढा।	The tastes are sixfold.	The juices are sixteen.	*	Does not understand non-standard form of धा pratyaya षोढा = sixfold and the sense of rasa as taste.
पदे द्विधा सुबन्तं तिङन्तं चेति।	There are two types of padams (words) - subanta and tinganta.	There are two types of verbs, the subjunctive and the triplet.	*	Does not understand vyakarana technical terms.
वनं पुलकितं सुन्दरम्।	The joyful forest is beautiful.	The forest is bright and beautiful.	?	Translation of पुलकित is inconsistent.
वनं पुलकितम्।	The forest is joyful.	The forest is thrilled.	?	Translation of पुलकित is inconsistent.
त्वं कुत्र असि?	Where are you?	Where are you?	Ok	
यूयं कुत्र स्थ?	Where are you (all)?	Where are you guys?	?	"Guys" seems too informal.
युवां कुत्र स्थः?	Where are you (two)?	Where are you guys?	?	Dual and plural information is not consistently differentiated.
भवान् कुत्र अस्ति?	Where are you?	Where are you at?	?	Where did "at" come from?
भवन्तौ कुत्र स्तः?	Where are you (two)?	Where are you two?	Ok	
भवन्तः कुत्र सन्ति?	Where are you (all)?	Where are you guys?	?	Dual and plural information is not consistently differentiated.
त्वं ग्रामे असि।	You are in the village.	You are in the village.	Ok	
युवां ग्रामे स्थः।	You (two) are in the village.	You are in the village.	?	Does not handle dual number and priority in case of mixed persons well.
यूयं ग्रामे स्थ।	You (all) are in the village.	You are in the village.	?	
भवान् ग्रामे अस्ति।	You are in the village.	You are in the village.	Ok	
भवन्तौ ग्रामे स्तः।	You (two) are in the village.	You two are in the village.	?	
भवन्तः ग्रामे सन्तिः।	You (all) are in the village.	You are in the village.	?	
गौः पर्णे खादति।	The cow eats (two) leaves.	The cow eats on the leaf.	?	Confuses neuter dual with saptami vibhakti. Cow eats on the leaf has no "yogyataa".

गौः द्वे पर्णे खादति।	The cow eats two leaves.	The cow eats two leaves.	Ok	Works ok if two is explicitly mentioned.
वृक्षात् पर्णे पततः।	(Two) leaves fall from the tree.	Leaves fall from the tree.	?	Does not distinguish dual from plural consistently.
वृक्षात् द्वे पर्णे पततः।	Two leaves fall from the tree.	Two leaves fall from the tree.	Ok	Works ok if two is explicitly mentioned.
यदि कयाचित् कवितया त्वधृदयम् अदारितं। तर्हि अहो बत! त्वया नैव काव्यम् अनुभूतम्॥	If your heart has not been torn apart by some poem, then oh my god! Poetry has not been experienced by you.	If any poem has torn your heart. Then oh my god! You have not experienced poetry.	*	Does not understand नञ् तत्पुरुष समास - अदारितम् = न दारितम्.
यदि कयाचित् कवितया त्वधृदयम् न दारितं। तर्हि अहो बत! त्वया नैव काव्यम् अनुभूतम्॥	If your heart has not been torn apart by some poem, then oh my god! Poetry has not been experienced by you.	If some poem has not torn your heart. Then oh my god! You have not experienced poetry.	Ok	

Part 2 of 2: English to Sanskrit

Test Input	Reference (expected) Output	GoogleTranslate Output	Rating	Remark
The chair is made of wood.	आसन्दं काष्ठेन निर्मितम्।	कुर्सी काष्ठेन निर्मिता भवति।	*	Uses Hindi/Urdu word कुर्सी for chair.
The table is made of wood.	उत्पीठिका काष्ठेन निर्मिता।	मेजः काष्ठेन निर्मितः अस्ति।	*	Uses Hindi/Urdu word मेज़ for table.
He did not come to work today as he is feeling a bit under the weather.	सः अद्य कार्यं कर्तुं न आगतः यतः सः किञ्चित् अस्वस्थः अस्ति।	सः अद्य कार्यं कर्तुं न आगतः यतः सः किञ्चित् मौसमस्य अधः अनुभवति।	*	Uses Hindi/Urdu word मौसम for weather. Does not understand the idiom "under the weather".
Armchair journalists on the internet are a dime a dozen.	आसन्दपत्रकाराः अन्तर्जाले बहवः।	अन्तर्जालस्य आर्मचेयरपत्रकाराः एकदर्जनं एकरूप्यकाणि भवन्ति।	*	Does not understand idioms such as "a dime a dozen".
Never were truer words spoken.	कदापि सत्यतरं वचनं न उक्तम्।	कदापि सत्यतरं वचनं न उक्तम् ।	Ok	
Truer words were never spoken.	सत्यतरं वचनं कदापि न उक्तम्।	सत्यतरं वचनं कदापि न उक्तम्।	Ok	
You are the apple of my eye.	त्वं मम नेत्रदीपः (महार्घः)।	त्वं मम नेत्रस्य सेबः असि।	*	Uses Hindi/Urdu word सेब for apple. Does not understand idiom "apple of the eye".
He has many servants at his beck and call.	तस्य सेवायै नैकाः सेवकाः सन्ति।	तस्य मुखे अनेके सेवकाः सन्ति।	?	Is मुखे the right translation of "beck and call"?
In order to develop an idea to its fullest, we must explore all facets of it.	कस्यचित् विचारस्य पूर्णविकासाय अस्माभिः तस्य सर्वाः पक्षाः अन्वेष्टव्याः।	विचारस्य पूर्णतया विकासाय अस्माभिः तस्य सर्वान् पक्षान् अन्वेष्टव्यम् ।	?	Should be सर्वे पक्षाः अन्वेष्टव्याः. Does not handle verb agreement for kartari/karmani correctly.
May the best team win.	उत्तमः दलः विजयं प्राप्नुयात्।	उत्तमं दलं विजयं प्राप्नुयात्।	?	दल in the sense of "team" is masc in Sanskrit
Thank you.	धन्यवादः।	धन्यवाद।	?	Should be धन्यवादः - Hindi influence?
Don't worry.	चिन्तामाऽस्तु।	चिंता मास्तु।	Ok	
How are you?	भवान् कथम् अस्ति?	भवान् कथमसि।	?	If भवान् is used, it should be अस्ति (प्रथमपुरुष)
I owe you.	अहं भवतः ऋणी अस्मि।	अहं भवतः ऋणी अस्मि।	Ok	Needs to assume male speaker, which is ok
She said, "I owe you".	"अहं भवत्याः ऋणिनी अस्मि" इति सा अवदत्।	सा अवदत्, "अहं भवतः ऋणी अस्मि" इति।	?	For female speaker being quoted, it should be ऋणिनी
She said that he was going home.	सः गृहं गच्छति इति सा अवदत्।	सा अवदत् यत् सः गृहं गच्छति।	Ok	
She said he was going home.	सः गृहं गच्छति इति सा अवदत्।	सा अवदत् यत् सः गृहं गच्छति।	Ok	
She said that he was going home yesterday.	सः ह्यः गृहं अगच्छत् इति सा अवदत्।	सा अवदत् यत् सः श्वः गृहं गच्छति।	*	श्वः is tomorrow, not yesterday! Seems to be Hindi influence, where the same word कल is used for both.

She said "I am going home".	"अहं गृहं गच्छामि" इति सा अवदत्।	सा अवदत् "अहं गृहं गच्छामि" इति।	Ok	
She said that she was going home.	अहं गृहं गच्छामि इति सा अवदत्।	सा गृहं गच्छामि इति अवदत्।	Ok	
She said she was going home.	अहं गृहं गच्छामि इति सा अवदत्।	सा गृहं गच्छति इति अवदत्।	?	Why the inconsistency with and without "that"?
Each time you open a book and read it a tree smiles knowing there is life after death.	यदायदापठनार्थं उद्घाटयसिपुस्तकम्। तदातदावृक्षस्स्मयते मरणोत्तरजीविज्ञः॥	प्रत्येकं पुस्तकं उद्घाट्य पठति तदा एकः वृक्षः मृत्योः अनन्तरं जीवनम् अस्ति इति ज्ञात्वा स्मितं करोति।	?	"Each time" should be "यदाकदाचित्"/"यदाकदापि" ("whenever")
Once the heart gets too heavy with pain, people don't cry. They just turn silent. They just turn silent completely.	वेदनया अभिभूतहृत्सु सत्सु नैव रुदन्ति जनाः। प्रत्युत निःशब्दा भवन्ति ते। केवलं पूर्णतया निःशब्दाः।	एकदा हृदयं वेदनायाः अतिभारं भवति तदा जनाः न रोदन्ति। ते केवलं मौनम् एव भवन्ति। ते केवलं सम्पूर्णतया मौनं कुर्वन्ति।	?	Here "once" should be "यदा" (when)
If a poem hasn't ripped apart your soul, you haven't experienced poetry.	यदि कयाचित् कवितया त्वधृदयम् न दारितं। तर्हि अहो बत! त्वया नैव काव्यम् अनुभूतम्॥	यदि कश्चन काव्यः भवतः आत्मानं न विदारितवान् तर्हि भवता काव्यस्य अनुभवः न कृतः।	Ok	
Ganesh and I are studying.	गणेशः च अहं च पठावः।	अहं च गणेशः पठामः।	*	Should be पठावः
I and Ganesh are studying.	अहं च गणेशः च पठावः।	अहं गणेशः च पठतः।	*	Should be पठावः
Harish, Ganesh and I are studying.	हरीशः च गणेशः च अहं च पठामः।	हरीशः, गणेशः, अहं च अध्ययनं कुर्मः।	Ok	
Harish, you and I are studying.	हरीशः च त्वं च अहं च पठामः।	हरीश, त्वं च अहं च अधीतवन्तौ।	?	Should be पठामः
You two and Ramesh are playing.	युवां च रमेशः च क्रीडथ।	भवन्तौ रमेशौ च क्रीडतः।	*	Should be क्रीडथ
Ramesh and you are playing.	रमेशः च त्वं च क्रीडथः।	रमेशः त्वं च क्रीडसि।	*	Should be क्रीडथः
Harish, Ramesh, you and I are playing.	हरीशः च रमेशः च त्वं च अहं च क्रीडामः।	हरीश रमेशः त्वं च अहं च क्रीडतः।	*	Should be क्रीडामः. Does not handle mixed-person dual/plural cases correctly or consistently.

Appendix C – The Systemic Gap example

Sanskrit original	Ideal Hindi translation	Google Hindi translation
1. अश्वे उपविशन् पिता पुत्रं पश्यति [aśve upaviśan pitā putraṃ paśyati]	1(HI). घोडे पर बैठा पिता अपने बेटे को देखता है	1(HG). पिता अपने बेटे को घोड़े पर बैठते हुए देखता है
2. पिता अश्वे उपविशन्तं पुत्रं पश्यति [pitā aśve upaviśantaṃ putraṃ paśyati]	2(HI). पिता घोडे पर बैठे अपने बेटे को देखता है	2(HG). पिता अपने बेटे को घोड़े पर बैठा देखता है

Sanskrit original	Ideal English translation	Google English translation
1. अश्वे उपविशन् पिता पुत्रं पश्यति [aśve upaviśan pitā putraṃ paśyati]	1(EI). The father seated on the horse sees his son	1(EG). The father looks at his son as he sits on the horse
2. पिता अश्वे उपविशन्तं पुत्रं पश्यति [pitā aśve upaviśantaṃ putraṃ paśyati]	2(EI). The father sees his son who is seated on the horse	2(EG). The father sees his son sitting on the horse

The difference between Sanskrit sentences 1 and 2 is relationship of the phrase meaning “seated on a horse” with the father in sentence 1 and with the son in sentence 2. This is marked by inflection agreement with the appropriate nouns, so both sentences 1 and 2 in Sanskrit are clear and unambiguous irrespective of the word order.

In Hindi too, ideally the phrase meaning “seated on a horse” should be in agreement with the appropriate nouns (as shown in the ideal translations 1HI and 2HI). However, Google’s Hindi translations (1HG and 2HG) do not show this agreement. Instead, 1HG effectively means “The father sees his son sitting on a horse” and 2HG effectively means “The father sees his son seated on a horse”. Both these sentences use an oblique case marker to make the phrase “seated on a horse” a preposition phrase rather than an adjective phrase of the respective nouns. Such usage is increasingly common in everyday Hindi.

As a consequence of the choice of syntax, both sentences suffer from the same prepositional attachment problem that exists in their English meaning, and both become ambiguous and can represent sentence 1 as well as 2. In the main paper, for ease of presentation, we have combined 1HG and 2HG into a single Hindi sentence that represents the way both sentences 1 and 2 would be typically translated in Hindi.