Final Draft Standard
on
Machine Translation Acceptance

Version 4.0

Ministry of Electronics & Information Technology
Government of India
Electronics Niketan, 6 CGO Complex
New Delhi 110003

Revision History Table

Sr. No.	Version	Date of Release	Pages Affected
1.	Initial Draft	20/07/2016	All
2.	Second version	22/09/2016	All
3.	Third Version	26/09/2016	8, 10, 11, 12, 14, 15
4.	Fourth Version	30/09/2016	All

Table of Contents

1. Background

Machine Translation is the process of automatically converting the text from one natural language into another natural language. Different organizations have been working for several years to build user acceptable machine translation systems and overcome the language barriers and enable easier communication.

1.1. Worldwide Activity of Machine Translation

A survey of the machine translation systems shows that, huge efforts are being taken to develop various machine translation systems worldwide. Many MT systems across the globe have already been developed for the most commonly used natural languages such as English, Russian, French, Japanese, Chinese, Spanish, Hindi and other Indian languages etc. SYSTRAN, LOGOS, METEO, Weidner and SPANAM are some of the well known results of the same. Some of the companies like CDAC, Google, Microsoft, are also offering Web based Machine Translation Services for limited sentences. In India domain specific MT systems for English to Indian Language and Indian Languages to Indian Languages have been developed and made available through TDIL Data Centre (www.tdil-dc.in).

1.2. Machine Assisted Translation

The availability of content in multiple languages has become one of the most significant aspects of communicating information between businesses, companies and their customers, organizations and countries. High level of accuracy and speed of producing translated content are important components of translation services. As per studies, a human translator can address at best 5-7 pages per day after which the translation loses its efficiency. As of now no machine translation system can perfectly translate the text into the target language, it can only aid the translation process. Hence, the term machine assisted/ aided translation (MAT) evolved.

MAT acts as a tool whereby translation efficiency may be increased. This is achieved by subsequently post-editing the output translation by the human. The time required to correct the errors in machine translated output is often seen as a measure of all post editing efforts.

Machine Assisted Translation is a powerful tool that has many purposes and can be used in a number of different ways:

- Quick translation of contents to understand the information in another language.

- Aided Tool/ Helping Hand for human translators for translating bulk documents.

- Instant translation of real time communication of chat, email or customer support communications.

1.3. Different Machine Translation Approaches

Worldwide, there are different technological approaches being used for the development of machine translation systems as briefed below:

i. Rule based MT

Rule-based machine translation is based on classical approach in which linguistic rules, grammar and bilingual dictionaries for source and target languages are used. It mainly covers the semantic, morphological, and syntactic regularities of each language.

ii. Example based MT

Example-based machine translation (EBMT) method mainly uses bilingual corpus with parallel texts as its main knowledge base. It uses case-based reasoning approach and it is a translation by analogy.

iii. Statistical MT

SMT a mathematical model, correspondences between the words in the source and the target language, which are learned from bilingual corpora.

iv. Tree Adjoining Grammar based MT

The Tree Adjoining Grammar (TAG) consists of a set of elementary trees, divided into initial and auxiliary trees. This works on tree-to-tree translation model and makes use of syntactic tree for both the source and target language. Parser and Generator modules recognize various grammatical entities in the English sentence, analyze them, represent them in different tree structures and synthesize equivalent Indian Language sentences on the basis of the derivational tree structure and the Transfer Grammar.

v. Analyze-Transfer-Generate based MT

The source languages text is preprocessed (collected, cleaned, and formatted) and analyzed. After language analysis, transfer of vocabulary and analyzed structure is carried out. And finally the target language text is generated.

vi. Hybrid & Pseudo Interlingua MT

This technology uses a pseudo Interlingua approach. It analyzes English only once and creates an intermediate structure with most of the

disambiguation performed. The intermediate structure is then converted to each target language through a process of text generation. It also uses rule based MT and Example based MT approaches.

1.4. Government of India Efforts in Machine Translation System Development

India has multiple languages and scripts, hence need of language translation is immense. Work in this area has been going on for several decades. Many premier academic and R&D organizations are engaged in the development of MT Systems for Indian languages.

Government of India has facilitated development of Machine Translation Systems for translation from English to Indian Languages and from Indian Languages to Indian Languages using different technological approaches. The language pairs addressed for English to IL are English-Hindi; Assamese; Bodo; Bangla; Gujarati; Malayalam; Marathi; Nepali; Punjabi; Oriya; Tamil; Telugu; and Urdu. Indian Language MT Systems include 9 bidirectional language pairs, i.e. Punjabi – Hindi –Punjabi, Telugu – Tamil – Telugu, Urdu – Hindi – Urdu, Hindi – Telugu – Hindi, Marathi – Hindi – Marathi, Bengali – Hindi – Bengali, Tamil – Hindi – Tamil, Kannada – Hindi – Kannada, Malayalam – Tamil – Malayalam.

2. History of MT Evaluation

In machine translation development, evaluation is very important, yet a difficult task. The difficulty arises due to some inherent characteristics of the language pairs, like simple word-level discrepancies to more structural variations for English to Indian languages, such as reduplication of words (चलते- चलते), free word order, etc. It is observed that the purpose of evaluation varies from stakeholder to stakeholder like it is different for Funding agency, MT developers and for MT end users or translators. Once the objective of evaluation is set in place, evaluations in conformity with the objectives need to be defined.

Translation quality is not an absolute concept, it needs to be assessed. Researchers have worked on MT evaluation techniques and evolved many subjective (Human) and automatic evaluation techniques.

2.1 Subjective (Human) Evaluation

In case of Indian languages, there is no single correct translation of a text, but multiple good translation options can exist. So comparing the translation with single reference does not give the actual quality of the MT Systems. In these cases subjective evaluation is more useful. Human evaluation allows measuring the quality of MT system over a set of users. Human evaluators can recognize

and weigh errors in translation correctly. Because of these strong arguments, subjective human evaluation is important. The main focus is on different types of subjective evaluation methods.
Following are the widely used subjective evaluation methods evolved for testing the outcomes of the Machine Translations Research and Development projects implemented under TDIL Programme, which focuses on the users and their needs.

i. 5-Point Scale for Accuracy Evaluation of Indian Languages MT.
ii. Modified 5-Point Scale for Comprehensibility & Fluency for Indian Languages
iii. 4-Point Scale for Accuracy provided by Google Translation

Limitations of Human Evaluation

Though human evaluation is very useful & informative; it has several inherent challenges & limitations which are:

i. Expensive & slow: Human evaluation is a time consuming activity and also labor-intensive, as while evaluating each sentence, evaluator needs to consider several quality criteria.

ii. Inter-evaluator agreement: Generally machine translation output is evaluated by multiple evaluators. Two evaluators do not give same score on the same data set, though the same evaluation guidelines/ training are provided. Perception about language, choice of words, may differ from human to human. So, in human evaluation, inter-annotator agreement issue always persists.

iii. Training to evaluators: Providing trainings and guidelines to the evaluators is very crucial and to clarify each scale in detail to the evaluators is a difficult task. Even after detailed training there is risk of wrong grading by the evaluators.

2.2 Automatic evaluation

Automatic evaluation is a faster, inexpensive and mostly language independent way of evaluating the translation quality. It is useful wherein frequent evaluations are required.

The quality of a translation is inherently subjective. Therefore, any metric must assign quality scores such that they correlate with human judgment of quality. Human judgment is the benchmark for assessing automatic metrics as humans are the end-users of any translation output. Even if a metric correlates well with human judgment in one study on one corpus, this successful correlation may not carry over to another corpus. Automatic evaluation metrics are BLEU,

NIST, METEOR and TER are more suitable for same language family pairs such as Latin based languages but these are not of much use for distant language pairs such as English to Indian Languages.

Limitations of Automatic Evaluation w.r.t. Indian Language MT systems

Due to the limitations of automatic evaluation, these are not directly applicable for Indian language machine translation evaluation.

i. Automatic measures e.g. BLEU, METEOR, NIST are not diagnostic as these are based on measures of string matching. These metrics do not provide feedback on the ability of MT systems to translate various aspects of the language.

ii. Word Order: These most popular metrics e.g. BLEU, NIST, PER, TER do not work well when used for evaluation of translation among distant language pairs like English to Indian languages. As English has a different word order than Indian languages. All the Indian languages need special attention to word order in the translation, otherwise meaning of the sentence changes completely, which often lead to incomprehensibility.

iii. Multiple correct translations: In case of BLEU, only exact match is considered and synonyms are not considered. All the Indian languages are morphologically rich; there can be multiple correct translations for single input sentence just word to word matching may lead to wrong results and make the evaluation process harder. Multiple correct translations may differ in word choice or the word order choice also.

iv. BLEU is N-gram precision based metric and does not care about the untranslated words in output translation. It gives poor co-relation with the human judgment.

v. No single automatic metrics can perform well for all the Indian language pairs. To work METEOR exceptionally well for all the language pairs, it needs huge corpus and a large repository of linguistic tools & resources.

Evaluation of MT system is important and there are various techniques for Indian languages MT evaluation, irrespective of MT development approaches. Both, human evaluation (subjective) and automatic evaluation (quantitative) have some advantages and limitations. In case of Indian languages, these methodologies are not very useful from end user's evaluation perspective. Hence, a simplified evaluation methodology for defining the criterion of accepting the machine translation output is required.

Details of various methods as listed above may be referred at Annexure-II.

3. Objective

This Standard (Code of Practice) document defines criterion for accepting the quality of translated content created by machine translation system. This standard will provide "minimum translation quality" required for machine translated content to industry and machine translation organizations.

This standard also defines acceptability metric in the form of various parameters, which measures the quality of the output of Machine Translation system based on post-editing efforts required to make the translated content acceptable by the end user. The metric will measure the acceptability of translation in terms of quality of output, irrespective of the source and target language and the machine translation techniques.

The evaluation procedure defined is simplified so that a person who knows both the source and target languages can easily evaluate the translated content.

4. Scope

This document is intended to serve as an acceptance criteria of machine translation output. It can be used by translation agencies, MT evaluation agencies and users to check the acceptability of machine translation output. Results based on evaluation of this metric will also be helpful to the MT researchers/ Developers for analyzing / improving the quality of Machine Translation System.

5. Machine Translation Acceptability

Machine translation acceptability is an inspection method that measures the system errors so that a decision could be taken about the usefulness of the machine translated content. The quality of machine translation output is judged on the time taken for post editing to achieve perfect translation versus time taken for manual translation. It is expected that use of machine translation should reduce the overall translation time to 50-60% minimum of the time taken over manual translation. Machine translation acceptance may also be defined as a criterion to check whether it is feasible to accept the MT output for post editing or to retranslate the entire content manually.

6. Machine Translation Acceptability Parameters

In order to judge the acceptability of MT output, various parameters are defined in simplified terminologies so that it can be used by any native language evaluators. Different grading scales are defined for each of the parameters to check the acceptance of machine translation output. Following are the parameters defined:-

6.1 Acceptability Parameters

i. Meaning Conveyed

The main objective of translation is to retain the information (meaning of the context) of source language in the target language(s). Therefore, it is important to check how much information from source language is conveyed in the target language. The following grades are defined to check this parameter.

0	No meaning at all or meaningless and hence, user is not able to guess the meaning of translated sentence
0.75	Word to word translation
1.5	Meaning of partial sentence conveyed
2	Meaning conveyed

ii. Sentence Structure

If the sentence structure of the output is not proper or the target language words are correct, but not as per the syntactic rules of target language then this error can be marked as sentence structure error. The following grades are defined to check this parameter.

0	Sentence structure is not at all proper and user prefers to re-translate the same sentence.
1	Sentence structure is partially correct and user prefers to do the post editing in the translated output.
2	Sentence structure is proper

iii. Word Inflection

Word inflection means the word is proper, but it requires a little modification. This word inflection error may contain error in gender translation, incorrect verb form, incorrect noun form, incorrect adjective form or adverb.

0	Word inflection error present in the translated sentence
1	No Word inflection error present in the translated sentence

iv. Spelling Mistakes

The spelling mistake is considered if any word in the translated sentence is misspelled. The following grades are defined to check this parameter.

0	The output translation has spelling mistake/s
1	The output translation has no spelling mistake

v. Suitability for Intended Purpose

This category verifies that whether the translated content satisfies the intended purpose, i.e. contextual choice, domain proximity, post-editing and user requirement.

0	Not suitable for intended purpose.
1	Suitable for intended purpose

vi. Transliteration:

Named Entity (NE) Transliteration:

Machine translation system should recognize named entities (proper names; including trade names, brand names, registered trademarks, place names, and personal names) in the source language text and should transliterate these NEs into target language. The following grades are defined to check this parameter.

Word Transliteration

Transliteration of source language words which are acceptable as it is in the target language. The following grades are defined to check this parameter.

0	Named Entity is not identified and transliterated/ Wrong word transliteration or where the transliteration is not desired in the target language domain.
1	Named Entity is identified and transliterated correctly Correct word transliteration, which are acceptable as it is in the target language

vii. Punctuations

Sometimes the punctuation marks used in source sentences are not be retained in translated sentences. Wrong placement of punctuation marks may change the entire meaning of the sentence.

0	Punctuation error [punctuation missing or wrongly placed]
1	Punctuation retained

viii. Numerals

"Numerals" category in the translated sentence may contain Numbers, Currency, Date, Time, Percentage, etc.

While translating the numbers, ordinals from source language to target language are not handled properly then it should be marked as "Numeral's Error" e.g. 2nd, 4th, 13th, 1980, 10 वी, are not translated properly, then should be marked as "Numeral's Error". The following grades are defined to check this parameter.

0	Numerals are not translated in the output translation
1	Numerals are correctly translated

ix. Abbreviations/Acronyms

Acronym is an abbreviation formed from the initial letters of a word. If some acronym is present in the source sentence, it needs to be checked in the translated content. If abbreviations/ acronyms are not generated properly, it will be marked as Acronym error. The following grades are defined to check this parameter.

0	Abbreviations/Acronyms not identified and translated properly
0.5	Abbreviations/Acronyms identified and translated correctly

x. Un-translated/ Extra Words

Untranslated Word - There might be a case like some of the words are not getting translated in target language and remain as it is (i.e. in the source language script) in the output sentence. User can mark this error as "Un-translated Words".

Extra Word - When extra words i.e. there is no corresponding reference word in the source sentence but these words are present in the output sentence.

The following grades are defined to check this parameter.

0	Untranslated Word/s OR Extra Words error is present
0.5	Untranslated Word/s OR Extra Words error is not present

6.2 Guidelines for Grading

- The evaluator should have access to the source language text while assigning the grades to the translated content.
- The evaluator will grade the above parameters for each sentence as per the defined grading scale.
- Once a grade is assigned to a word in a sentence for any error, another grade cannot be assigned for the same word in that sentence. This is to avoid wrong grading i.e. if a wrong word inflection is observed in the output sentence, then it should be marked as word inflection error only and not gain as a spelling error.
- Evaluation should be conducted on a data set of minimum 100 sentences covering different grammatical structures and various parameters listed above; however, all the above mentioned parameters may not be applicable for each sentence. The sentences selected for evaluation should be of minimum six words each and from the intended domain.
- It is suggested that minimum 03 evaluators should be engaged to reduce the subjectivity in the evaluation.
- Training to the evaluators is a very important component of evaluation and it should cover the different post editing features and grades as well.

6.3 Acceptability Score Calculation

S.No.	Acceptability Parameters	Grade Assigned	Weigh t	Maximum/ Minimum Score
i.	Meaning Expressed	2/1.5/0.75/0	20	40/30/15/0 0
ii.	Sentence structure	2/1/0	10	20/10/0
iii.	Word Inflection Error	1/0	5	5/0
iv.	Spelling Mistakes	1/0	5	5/0
v.	Suitable for Intended Purpose	1/0	10	10/0
vi.	Transliteration	1/0	5	5/0
vii.	Punctuations	1/0	5	5/0
viii.	Numerals	1/0	5	5/0
ix.	Abbreviations/Acronyms	0.5/0	5	2.5/0
x.	Extra Word/Un-translated Words	0.5/0	5	2.5/0

(A) Formula for calculating Acceptability score for single evaluator (E_i)

$$E_i = [(X_1 + X_2 + X_3 + \ldots\ldots + X_n)/n]$$

Where n = Total number of sentences (Minimum 100 sentences)
Xi = Sum of scores given by evaluator for the applicable parameters in each sentence.

(B) Final Acceptability score (average of evaluator's score) is calculated with this formula:-

$$\text{Acceptability Score} = [(E_1 + E_2 + E_3 + \ldots\ldots + E_N)/N]$$

Where N = Total number of evaluators

6.4 Recommended Acceptability Score

A Machine Translation Output should get minimum *Acceptability Score* of 50 based on above recommended formula. This score indicates that the Machine Translation output is acceptable for post editing and making it useful for the intended purpose.

The *Acceptability Score* should be calculated on a minimum set of 100 sentences and 3 evaluators however more number of sentences and evaluators may be taken for better evaluation.

Annexure I: Software Engineering Advisory Parameters

Along with the machine translation output acceptance parameters, this document also defines machine translation software advisory parameters. These parameters will relate mainly to the MT System.

a. Translation Production Speed

In case of bulk job translations the performance speed of the MT Engine is very crucial. The machine translation software shall be evaluated for the translation speed at character/word & sentence level. The total translation time for the text or file shall be measured and speed shall be calculated in character/second or word/second etc.

b. Portability

It is observed that different operating systems and devices like Mobile, Tablet, Laptop have different script processing/ input methods. Hence, the functionality of the MT Software shall be tested for different devices and Operating Systems to ensure inter-operability.

c. Usability

The look and feel shall be evaluated from the users' point of view.

- Menu Navigation: Menus shall be readable, self explanatory and there shall be no broken links.
- Help Availability: On-line help shall be available to the user with suitable examples.
- User Documentation: There should be a comprehensive user manual describing all the functions along with examples. It should be clear, understandable and properly indexed.

Annexure II: Details of Earlier Human Evaluation Methodologies

In case of Indian languages, there is no single correct translation, but multiple good translation options can exist. So instead of comparing the translation with single reference, subjective evaluation is more useful. Human evaluation allows measuring the quality of MT system by a group of end users. Translations are produced for end users, hence end users are the right measure of quality of translation and end users can recognize and weigh errors in translation correctly. Because of these strong arguments, subjective human evaluation is important.

In this section, we give an overview of the most well known evaluation methods which focuses on the users and their needs. The main focus is on different types of subjective evaluation methods. Following are the grading systems used for evaluation of MT systems. Initially, the 7-Point Russian Grading System was proposed then 5-point scale for Accuracy was proposed and used. Afterwards, 5-Point Scale: Comprehensibility & Fluency were used.

1. 5-Point Scale: Accuracy (Proposed Scale for Indian Languages)

Various MT Systems have been developed under the TDIL-DeitY funded MT projects. A 5-point evaluation scale was evolved to assess their performance. The approach was formulated with stress on Sprachgefühl i.e. focus on usability and the native speaker's expectations and the translation quality is measured in terms of comprehensibility of output.

Grade 0	No output provided by the engine concerned.
Grade 1	The translated output is not comprehensible.
Grade 2	Comprehensible after accessing the source text.
Grade 3	Comprehensible with difficulty.
Grade 4	Acceptable since the text is comprehensible.

2. Modified 5-Point Scale: Comprehensibility & Fluency

5-Point scale for accuracy mentioned above has Grade 0 for "no output provided by system". This was mainly the result of MT Engine system failure. Experts deliberated on the issue and grade -1 was evolved for No Output by system due to technical reasons such as buffer clearance, etc. Hence, a Modified 5-point scale was evolved as presented below. In modified, where performance of the MT system is measured on two parameters (1) Comprehensibility & (2) Fluency.

Grade -1	No Output OR buffer clearance issue
Grade 0	Nonsense (If the sentence doesn't make any sense at all – it is like someone speaking to you in a language you don't know)
Grade 1	Some parts make sense but is not comprehensible over all (e.g., listening to a language which has lots of borrowed words from your language – you understand those words but nothing more)
Grade 2	Comprehensible but has quite a few errors (e.g., someone who can speak your language but would make lots of errors. However, you can make sense out of what is being said)
Grade 3	Comprehensible, occasional errors (e.g., someone speaking Hindi getting all its genders wrong)
Grade 4	Perfect (e.g., someone who knows the language)

3. 4-Point Scale for Accuracy by Google Translation

Following is the 4-point scale defined by Google Translation for evaluation of English→French.

Grade	Scale	Description
Grade 0	Poor	None of the content is translated well
Grade 1	Fair	Only some of the content is translated well
Grade 2	Good	Most of the content is translated well
Grade 3	Excellent	All the content is translated well

Annexure III: GLOSSARY

Adequacy	Adequacy refers to the degree to which information present in the source text is communicated in target translation. The objective of the adequacy is to determine the extent to which all of the content of a text is conveyed, regardless of the quality of the language in the candidate translation.
BLEU	BLEU stands for Bilingual evaluation under study. It compares the translated machine output with the reference output generated by professional human translator. BLEU is precision based and it uses modified N gram precision. Word precision account for adequacy and n gram precision for n =1, 2, 3 account for fluency
Blind testing	In blind testing, evaluators have no access to the source text, this is to eliminate bias scoring.
Comprehensibility	Comprehensibility is a measure of how easy is a text to understand.
Fidelity	Fidelity is measurement of correctness of the information transferred from source language to the target language. It is a subjective evaluation of the measure in which the information contained in the sentence of the original text reappears without distortion in the translation.
Fluency	Fluency refers to the degree to which the target is well formed according to the rules of the target language. The objective of fluency evaluation is to determine how much like "good fluent" a translation appears to be without taking into account the correctness of information.
Intelligibility	Intelligibility is a measure of how "understandable" the sentence is. Intelligibility is measured without reference to the original source sentence.
METEOR	This metric is based on word to Word alignment between candidate and reference translation. The final Meteor score is between 0 and 1 which is a harmonic mean of unigram precision and unigram recall.
NIST	The NIST metric is based on the BLEU metric. Doddington established NIST in 2002 which is similar to BLEU except, it assigns a weight to each unigram depending upon its uniqueness or how informative the n-gram is
Open testing	Open testing, evaluators will have access to source text while evaluating MT translation output.
TER	TER (Translate Error Rate) was proposed by Snover and Dorr 2006, is a more. It represents the number of edits necessarily required to transform the machine output to reference translation.

References:

1. Evaluation of machine translation [Online]. Available: http://en.wikipedia.org/wiki/Evaluation_of_machine_translation.[January 9, 2014]
2. Sangal Rajeev et al. Machine Translation System:Shakti[Online]. Available: http://gdit.iiit.net/~mt/shakti/ , 2003.[January 20,2014]
3. Word error rate [Online]. Available: http://en.wikipedia.org/wiki/Word_error_rate. [January 28,2014]
4. Martin Thoma. Word Error Rate Calculation [Online]. Available: http://martin-thoma.com/word-error-rate-calculation. 2013. [February 4, 2014]
5. Sara Stymne, Machine Translation Evaluation [Online].Available: http://www.ida.liu.se/labs/nlplab/gslt/mt-course/mteval-sarst.pdf. [February 10,2014]
6. Evaluation of machine translation [Online]. Available: http://www.translationdirectory.com/articles/article1814.php . 2008. [February 27,2014]
7. Goyal, Vishal, and Gurpreet Singh Lehal. "Evaluation of Hindi to Punjabi machine translation system." arXiv preprint arXiv:0910.1868 (2009).
8. Josan, Gurpreet Singh, and Gurpreet Singh Lehal. "A Punjabi to Hindi machine translation system." 22nd International Conference on on Computational Linguistics: Demonstration Papers. Association for Computational Linguistics, 2008.
9. Sinha, R. M. K., and A. Jain. "AnglaHindi: an English to Hindi machine-aided translation system." MT Summit IX, New Orleans, USA (2003): 494-497.
10. Balyan, Renu, et al. "A diagnostic evaluation approach for english to hindi MT using linguistic checkpoints and error rates." Computational Linguistics and Intelligent Text Processing. Springer Berlin Heidelberg, 2013. 285-296.
11. Joshi, Nisheeth, et al. "HEVAL: Yet Another Human Evaluation Metric." arXiv preprint arXiv:1311.3961 (2013).
12. Zhou, Ming, et al. "Diagnostic evaluation of machine translation systems using automatically constructed linguistic check-points." Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1. Association for Computational Linguistics, 2008.
13. Specia, Lucia, et al. "Predicting machine translation adequacy." Machine Translation Summit. Vol. 13. No. 2011. 2011.
14. Correa, Nelson. "A fine-grained evaluation Framework for machine translation system development." MT Summit IX. 2003.
15. Van Slype, Georges. "Critical study of methods for evaluating the quality of machine translation." Prepared for the Commission of European Communities Directorate General Scientific and Technical Information and Information Management. Report BR 19142 (1979).
16. Goyal, Vishal. Development of a Hindi to Punjabi Machine Translation system. Diss. Ph. D. Thesis, Punjabi University, Patiala, 2010.

17. Kalyani, Aditi, et al. "Assessing the Quality of MT Systems for Hindi to English Translation." arXiv preprint arXiv:1404.3992 2014.
18. Vilar, David, et al. "Error analysis of statistical machine translation output."Proceedings of LREC. 2006.
19. Stymne, Sara, and Lars Ahrenberg. "On the practice of error analysis for machine translation evaluation." LREC. 2012.
20. Ananthakrishnan, R., et al. "Some issues in automatic evaluation of english-hindi mt: more blues for bleu." ICON 2007.
21. Joshi, Nisheeth, Hemant Darbari, and Iti Mathur. "Human and Automatic Evaluation of English to Hindi Machine Translation Systems." Advances in Computer Science, Engineering & Applications. Springer Berlin Heidelberg, 2012. 423-432.
22. Doherty, Stephen, Sharon O'Brien, and Michael Carl. "Eye tracking as an MT evaluation technique." Machine translation 24.1 (2010): 1-13.
23. Vaishali Gupta, Nisheeth Joshi and Iti Mathur. "Subjective and Objective Evaluation of English to Urdu Machine Translation." 1310.0578.pdf