Towards human linguistic machine translation evaluation

Marta R. Costa-jussà

Institute for Infocomm Research. 1 Fusionopolis Way, 21-01 Connexis (South Tower) Singapore 138632 E-mail: vismrc@i2r.a-star.edu.sg

Mireia Farrús

Pompeu Fabra University. C/ Tanger/Roc Boronat, 08018 Barcelona. E-mail:mfarrus@upf.edu

Abstract

When evaluating machine translation outputs, linguistics is usually taken into account implicitly. Annotators have to decide whether a sentence is better than another or not, using, for example, adequacy and fluency criteria or, as recently proposed, editing the translation output so that it has the same meaning as a reference translation, and it is *understandable*. Therefore, the important fields of linguistics of meaning (semantics) and grammar (syntax) are indirectly considered.

In this study, we propose to go one step further towards a linguistic human evaluation. The idea is to introduce linguistics implicitly by formulating precise guidelines. These guidelines strictly mark the difference between the sub-fields of linguistics such as: morphology, syntax, semantics, and orthography. We show our guidelines have a high inter-annotation agreement and wide error coverage. Additionally, we examine how the linguistic human evaluation data correlate with: among different types of machine translation systems (rule and statistical-based); and with adequacy and fluency.

Keywords: Machine translation, linguistic evaluation

1. Introduction

Evaluation in machine translation is a challenging task. As a consequence of the increased interest in enhancing machine translation systems, there is a correspondent interest in improving machine translation evaluation.

Evaluation of a translation output is not an easy task even for human beings because translation involves different types of knowledge, such as linguistic and cultural. Different translators may have different criteria. However, human judgments of performance have been the gold standard of MT evaluation metrics.

There have been several proposals for human evaluation which have been widely used by the scientific community. Measuring in *adequacy* and *fluency* was proposed by [17] and it is still a standard evaluation criteria. *Adequacy* is a rating of how much information is transferred between the source and the target language, and *fluency* is a rating of how good the target language is.

The most recent human evaluation approach that was chosen as the official machine translation evaluation metric for DARPA's Global Autonomous Language Exploitation (GALE) program [12] was HTER (Human-targeted Translation Edit Rate). HTER involves a procedure for creating targeted translations. Annotators compare the translation output against a reference translation, and they modify the output so that it has the same meaning as the reference, and is *understandable*. Each inserted/deleted/modified word or punctuation mark counts as one edit, while shifting a string of any number of words, by any distance, counts as one edit [13].

Other works such as [16] propose a 5-category schema that does not use linguistic criteria. The errors are classified in five big classes: *incorrect words*, *missing words*, *word order*, *unknown words* and *punctuation*. Flanagan classification [6] lists a series of errors that are language pair dependent. The author classifies the errors in 19 different categories for the English-to-French translation, plus three more categories to be added in the English-to-German translation. Evaluations of different MT systems for a range of linguistic checkpoints have been carried out for English-Chinese [18], Italian-English, German-English and Dutch-English [11].

As far as we are concerned, the above evaluations (except for *adequacy* and *fluency*) do not report an inter-annotation agreement study. In any case, there has not been a formal proposal of linguistic evaluation guidelines for machine translation. The main advantages of a linguistic evaluation would be:

- Propose precise linguistic guidelines that allow for a high inter-annotation agreement.
- Provide a linguistic classification of the translation output errors.
- Provide new information to enhance the machine translation systems.
- Evaluation is done without a reference.

The main drawbacks of such an evaluation would be that it requires bilingual annotators and it would be time consuming. However, nowadays we can take advantage of crowd-sourcing platforms (such as Amazon's Mechanical Turk[1], Crowdflower[2]) to reduce these types of drawbacks. Crowd-sourcing enables requesters to tap from a global pool of non-experts to obtain rapid and affordable answers to simple Human Intelligence Tasks (HITs), which can be subsequently

[1] `https://www.mturk.com`
[2] `http://crowdflower.com/`

used to train data-driven applications. A number of recent papers on this subject point out that non-expert annotations, if produced in a sufficient quantity, can rival and even surpass the quality of expert annotations, often at a much lower cost [14], [15]. However, this possible increase in quality depends on the task at hand and on an adequate HIT design [7], which motivates the creation of detailed guidelines.

The rest of the paper is organized as follows. Next section briefly describes the linguistic guidelines. Section 3 reports the experimental results with these linguistic guidelines. Particularly, we exploit the linguistic guidelines to show correlation results at the segment level between linguistic evaluations and different types of systems. Additionally, we test the linguistic guidelines by computing the correlation with *adequacy* and *fluency* results. Finally, section 4 discusses most relevant conclusions.

2. Linguistic guidelines

We consider that linguistic guidelines for a machine translation system should be specific for the target language. However, they may be generalizable for different source languages. In this case, we are using guidelines specific for the Catalan language. The guidelines consider four relevant linguistic evaluations: orthographic (language writing standardization); morphological (internal structures of words and how they can be modified); semantic (meaning of individual words and combinations, and how these form the meanings of sentences); and syntactic (word combination to form grammatical sentences). The guidelines should classify any error committed by a translation system into one of these categories.

The linguistic guidelines have been designed for the Catalan target language using the translation output of the Universitat Politècnica de Catalunya (UPC) statistical machine translation system [9] over a Spanish-to-Catalan test set[3]. The guidelines were designed by a Catalan linguist. Next, the annotation guidelines are summarized.

- **Orthographic errors** include punctuation marks, erroneous accents, letter capitalization, joined words, spare blanks coming from a wrong detokenisation, apostrophes, conjunctions and errors in foreign words.

    1. Punctuation marks. Include a wrong use, missing punctuation and extra punctuation (exclamation and interrogation marks, full stops, commas, colons, semicolons, dots, etc.). E.g.
    Source: Es factible, pero hay que tener en cuenta tres obviedades:
    Target: És factible, ara cal tenir en compte tres obvietats.

[3]This test set was of 711 sentences (around 16k words) extracted from *El País* and *La Vanguardia* newspapers [4]

2. Accents. Include accented vowels when not necessary, missing accents and erroneous accents. E.g.
   Source: La llegada de Obama y la situación interna del régimen islamista deparan una oportunidad.
   *Target: *L'arribada d'Obama i la situacio* interna del règim islamista ofereixen una oportunitat.*
   Correct target: L'arribada d'Obama i la situació interna del règim islamista ofereixen una oportunitat.
3. Capital and lower case letters. This refers to wrong capital letters within a sentence, lower case letters at the beginning of a sentence, and lower case letters in acronyms or proper nouns. E.g.
   Source: El enorme peligro de este camino sería privar a un régimen aislado y teóricamente revolucionario del enemigo supremo.
   Target: L'enorme perill d'aquest camí seria privar a un règim aïllat i teòricament revolucionari de l'enemic Suprem.
4. Joined words. This is a less common error committed where two consecutive words are erroneously joined. E.g.
   Source: Pero, aun siendo funcional para resolver y a la vista de los resultados (...)
   *Target: *Però, fins i tot sent funcional per resoldre ia la vista dels resultats.*
   Correct target: Però, fins i tot sent funcional per resoldre i a la vista dels resultats.
5. Extra spaces. This error is usually committed due to non-detokenising when required or detokenising into the wrong direction. E.g.
   Source: "hola"
   Target: " hola "
6. Apostrophe. Apostrophe is commonly used in Catalan to elide a sound. In some cases, some of the words that should be apostrophofised are not apostrophised (missing apostrophe) and viceversa (extra apostrophe). E.g.
   Source: Sólo hace 25 años que sabemos la historia que se oculta tras esa imagen turbadora
   *Target: *Només fa 25 anys que sabem la història que se amaga darrere aquesta imatge torbadora.*
   Correct target: Només fa 25 anys que sabem la història que s'amaga darrere aquesta imatge torbadora.

- **Morphological errors** include lack of gender and number concordance, apocopes, errors in verbal morphology (inflection) and lexical morphology (derivation and compounding), and morphosyntactic changes due to changes in syntactic structures.

  1. Lack of gender concordance. Some words are given a different gender in different languages. For instance, the word smile is feminine in Spanish (la sonrisa) and masculine in Catalan (el somriure). It is

then common to find a lack of gender concordance in articles and adjectives with a noun that changes its gender from one language to the other, especially in statistical systems, where there are no rules to solve it. E g.
Source: El balón llegó tarde
*Target: *El pilota va arribar tard.*
Correct target: La pilota va arribar tard.

2. Lack of number concordance. Although it is less common, some words are given a different number in different languages. For instance, the word money is singular in Spanish (el dinero) and plural in Catalan (els diners). Like in the previous case, this causes a lack of number concordance in articles and adjectives with the consecutive noun, especially also in statistical systems, where there are no rules to solve it. E. g
Source: El gobierno se ha gastado todo el dinero de los ciudadanos.
*Target: *El govern s'ha gastat tot el diners dels ciutadans.*
Correct target: El govern s'ha gastat tots el diners dels ciutadans.
3. Verbal morphology. This error refers to a verb that is not correctly inflected, a common error in a very inflected language such as Catalan. The most common cases are the translation of an inflected verb into the infinitive form, or the lack of person concordance.
Source: El mismo que usted puede ahora constatar en la exposición Bacon.
*Target: *El mateix que vostè pugues ara constatar en l'exposició Bacon.*
Correct target: El mateix que vostè pot ara constatar en l'exposició Bacon.
4. Lexical morphology It concerns basically word formation: derivation and compounding, like the use of a derivate in a wrong way (e.g. *lliguer* instead of *de la Lliga*) or a wrong compounding (e.g. *històric-social* instead of *historicosocial*).

- **Semantic errors** include no correspondence between source and target words, non-translated but necessary source words, missing target words, and non-translated proper nouns or translated when not necessary. Additionally, includes polysemy, homonym, and expressions used in a different way in the source and target languages.

    1. Polysemy. A polyseme is a word with multiple meaning, which shares the same origin. A polysemic error occurs when the incorrect meaning is chosen in the target language. E.g. the Catalan conjunction *perquè* has two different meanings: *porque (because)* and *para que (in order to)*, which causes usually translation errors.
    2. Homonymy. Homonymy is found when two or more words share the same spelling and the same pronunciation but have different meanings, usually as a result of having different origins. Like polysemes,

homonym errors occur when the incorrect meaning is chosen in the target language. E.g. the Spanish adverb *solo*, which can also be and adjective, is translated by the Catalan adjective *sol* instead of the corresponding adverb *només*.

3. Incorrect word. This error is detected when there is no correspondence at all between the source word and the translated target word. It is normally found in statistical systems, where the word is translated incorrectly mainly due to alignment errors. E.g.
   Source: No llegaron hasta las cuatro de la tarde.
   *Target: *No van arribar fins a les quatre.*
   Correct target: S'arreglen més sabates que mai,
4. Unknown word. This refers to a non-translated source word, which is left intact in the target side. E.g.
   Source: el caso Dutroux, que convulsionó a Bélgica a principios de los noventa,
   *Target: *el cas Dutroux, que convulsionó Bèlgica a principis dels noranta*
   Correct target: el cas Dutroux, que va convulsionar Bèlgica a principis dels noranta
5. Missing target word This refers to a non-translated source word, which is missing in the target side. E.g.
   Source: Fue una decisión impopular, pero seguramente justa.
   Target: Va ser una decisió impopular, segurament justa.
6. Proper nouns This error concerns non-translated proper nouns (i.e. unknown proper noun) or translated when not necessary (for instance, not being detected as proper noun but as common noun to be translated). E.g.
   Source: Zapatero se negó.
   *Target: *El sabater s'hi va negar.*
   Correct target: Zapatero s'hi va negar.

- **Syntactic errors** include errors in prepositions, errors in relative clauses, verbal periphrasis, clitics, missing or spare article in front of proper nouns, and syntactic element reordering.

1. Prepositions This error refers to prepositions not elided in the target language (extra prepositions), prepositions not inserted in the target language (missing prepositions), or source prepositions maintained in the target language instead of a new correct target preposition (incorrect prepositions). E.g.
   Source: Debería ser recusado en favor de otro juez.
   *Target: *Hauria de ser recusat en favor d'un altre jutge.*
   Correct target: Hauria de ser recusat a favor d'un altre jutge.
2. Relative pronouns. Due to its syntactic complexity, the use of relative clauses involving relative pronouns refering to previous elements usually leads to erroneous translations. E.g.

Source: Murieron tres personas al colisionar un Ford Scort con un Renault Scenic cuyo conductor sufrió heridas leves.
*Target: *Hi van morir tres persones al topar un Ford Scort amb un Renault Scénic amb un conductor va patir ferides lleus.*
Correct target: Hi van morir tres persones en topar un Ford Scort amb un Renault Scénic el conductor del qual va patir ferides lleus.

3. Verbal periphrasis. The use of verbal periphrasis, especially when they involve prepositions that differ in the different languages, usually leads to translation errors, as well (e.g. the Spanish verbal periphrasis *tener que (have to)* is usually translated literally into Catalan as *tenir que* instead of the correct periphrasis *haver de*).
4. Clitics Include an incorrect syntactic function of the pronoun or a wrong clitic-verb combination. E.g.
Source: El niño se cayó por las escaleras de su casa.
*Target: *El nen es va caure per les escales de casa seva.*
Correct target: El nen va caure per les escales de casa seva.
5. Articles. This error refers to missing or extra articles in front of proper nouns. E.g.
Source: Rosa entró en el despacho del dueo
*Target: *Rosa va entrar al despatx del propietari*
Correct target: La Rosa va entrar al despatx del propietari
6. Reordering. It refers to a syntactic reordering of the elements of the sentence.

A list of the linguistic errors can be found in Table 1.

Orthographic	Morphologic	Semantic	Syntactic
Puntuation marks	Gender concordance	Polysemy	Prepositions
Accents	Number concordance	Homonymy	Relative pronouns
Capital and lower case letters	Verbal morphology	Incorrect words	Verbal periphrasis
Joined words	Lexical morphology	Unknown words	Clitics
Extra spaces		Missing target word	Articles
Apostrophe		Proper nouns	Reordering

Table 1: *Guidelines summary.*

3. Experiments

This section describes the experiments that were designed to evaluate the performance of the linguistic guidelines briefly reported in the previous section. First, we wanted to evaluate the inter-annotation agreement. Second, we wanted to test the coverage of the linguistic errors and the generalization to a difference source language. Finally, we compute the correlation of the linguistic evaluations: among different translation systems, and with standard human evaluation methods such as *adequacy* and *fluency*.

3.1. Data set

The test corpus falls within the medicine domain. This medical corpus was kindly provided by the UniversalDoctor project, which focuses on facilitating communication between health-care providers and patients from various origins [4]. Table 2 summarizes the number of sentences, words and vocabulary of the medical corpus.

	English
Sentences	630
Words	4073
Vocabulary	1050

Table 2: *Corpus statistics of the English medical test set.*

3.2. Machine translation systems

As translation systems we used 4 freely available systems in the web. They include two rule-based MT (RBMT) systems, Apertium and Translendium, and two statistical MT (SMT) systems, Google Translate and UPC. All systems are used with their respective versions date of *1st of February 2010*.

- **Apertium** platform[5] is an open-source RBMT system originally based on existing translation systems that have been designed by the Transducens group at the Universitat d'Alacant (UA). The system uses a shallow-transfer machine translation technology.

- **Translendium**[6] is developed by Translendium S.L., a Catalan company located in Barcelona and subsidiary of the European group *Lucy Software*, made up of linguists and computer scientists with more than fifteen years of experience in the machine translation field. The translation engine consists of a modular structure of computational grammars and lexicons that makes possible to carry out a morphosyntactic analysis of the source text and then transfers it into the target language.

- **Google Translate**[7] is a SMT system developed by Google's research group for more than 50 languages. The system uses billions of words of text, both monolingual text in the target language. Google is constantly working to support more languages and introduce them as soon as the automatic translation meets their standards.

[4] *http://www.universaldoctor.com*
[5] *http://www.apertium.org/*
[6] *http://www.translendium.com*
[7] *http://translate.google.com/*

- **UPC** system[8] is developed at the Universitat Politècnica de Catalunya. Based on a Ngram translation model integrated in an optimized log-linear combination of additional features, it is mainly a statistical system, although it also includes additional linguistic rules to solve some errors caused by the statistical translation [4].

3.3. Inter-annotation agreement in adequacy and fluency human evaluation

The evaluation in *adequacy* and *fluency* was performed by three annotators Catalan native and fluent in English. The rank of *adequacy* and *fluency* was from 1 (good) to 5 (bad). All annotators evaluated 2520 (630*4) sentences both in *adequacy* and *fluency*. The inter-annotation agreement was evaluated with the weighted kappa [2] using a quadratic distance between errors. The weighted kappa was 0.62 which is qualified as 'good' according to [8].

3.4. Inter-annotation agreement in the linguistic human evaluation

The linguistic evaluation was performed by three annotators Catalan native and fluent in English. The errors are reported according to the following linguistic evaluations: orthographic, morphological, semantic and syntactic, as described in section 2.

Annotators were not able to find one single error that was not reported in the guidelines. This was one of the main objectives of the guidelines and it is a great achievement because the guidelines were designed on a different set from the test set with a different source language. This means that these guidelines designed for a particular target language may be used for different languages pair with common target.

We evaluated the inter-annotation agreement with the weighted kappa (k) [2] using a linear unitary distance between errors.

$$k = 1 - \frac{\sum_{i=i}^{k} \sum_{j=i}^{k} w_{ij} x_{ij}}{\sum_{i=i}^{k} \sum_{j=i}^{k} w_{ij} m_{ij}}$$

where k is the number of codes (in our case four categories) and w_{ij}, x_{ij}, m_{ij} are elements in the weight, observed and expected matrices, respectively. The weighted kappa was 0.75 which is good according to [8]. This kappa is quite high when comparing it to other inter-annotation kappas in MT evaluation [1] and it is due to the accurate design of the linguistic guidelines.

To sum up, we are boosting kappa by giving strict guidelines, which is different from relying on the holistic evaluation that provides the adequacy and fluency criteria. Depending on the application, we would prefer one evaluation or the other.

[8] *http://www.n-ii.org/*

3.5. Adequacy and fluency results

Table 3 shows the results of the translation evaluation from the 4 different system outputs. Notice that Google is ranked the best system in adequacy and Translendium is ranked the best system in fluency.

English-to-Catalan	Adequacy	Fluency
Apertium	2.9	2.5
Google	4.1	3.8
Translendium	3.9	4.0
UPC	2.8	2.0

Table 3: Adequacy and fluency results for English-to-Catalan translation output.

3.6. Linguistic human evaluation results

Table 4 shows the results of the translation evaluation from the 4 different system outputs. Notice that the semantic errors are the more common, and the orthographic errors are the less common. If we rank systems by orthography, Apertium is the best system. If we rank systems by morphology or syntax, Translendium is the best one. And if we rank systems by semantics, Google is the best one. Therefore, this evaluation may be worth to decide which system is better for a specific application. For example, if Ione requires tourist information, one may be only interested in the meaning of the translation, in this sense one may choose Google, which has the lowest number of semantic errors.

English-to-Catalan	Sent. w/errors	Total errors	Ort.	Mor.	Sem.	Syn.
Apertium	464	731	10	79	463	179
Google	305	492	27	72	232	161
Translendium	324	478	31	30	293	124
UPC	519	1168	33	139	715	281

Table 4: Linguistic evaluation results for English-to-Catalan translation outputs: number and type of linguistic errors.

Previous experiments with these guidelines can be found in [4], [3] and [5].
x

3.7. Correlation between linguistic evaluations and adequacy and fluency

We performed the correlation at the level of segment between the linguistic judgments and the *adequacy* and *fluency* criteria.

We performed the correlation at the level of segment using the Kendall's τ_B rank correlation among the different linguistic evaluations and systems. Let $(x_1, y_1), (x_2, y_2), ...(x_n, y_n)$ be a set of joint observations from two random variables X and Y respectively (f. e. orthography and semantics). Any pair of observations (x_i, y_i) and (x_j, y_j) are said to be concordant if the ranks for both elements agree: that is, if both $x_i > x_j$ and $y_i > y_j$ or if both $x_i < x_j$ and $y_i < y_j$. They are said to be discordant, if $x_i > x_j$ and $y_i < y_j$ or if $x_i < x_j$ and $y_i > y_j$. If $x_i = x_j$ or $y_i = y_j$, the pair is neither concordant, nor discordant. Given that presumably we've got many ties, we use the Kendall τ_B coefficient which makes adjustments for ties and it is defined as:

$$\tau_B = \frac{(number\ of\ concordant\ pairs) - (number\ of\ discordant\ pairs)}{\sqrt{(n_0 - n_1)(n_0 - n_2)}}$$

where a concordant pair is a pair of two translations of the same segment in which the rank given by the number of errors calculated from the corresponding linguistic level agree; in a discordant pair, they disagree. Ties are adjusted as shown in the denominator:

$$n_0 = n(n-1)/2\,; n_1 = \sum_i t_i(t_i - 1)/2\,; n_2 = \sum_j u_j(u_j - 1)/2$$

where n is the total number of pairs, t_i is the number of tied values in the i^{th} group of ties for the first quantity and u_j is the number of tied values in the j^{th} group of ties for the first quantity. The possible values of τ_B range between 1 (where all pairs are concordant) and -1 (where all pairs are discordant). Thus the higher the value for τ_B the more similar the linguistic evaluations. When τ_B is zero, it means linguistic evaluations are independent. In all cases, correlations followed a statistically significant trend [10].

Here, a concordant pair is a pair of two translations of the same segment in which the ranks calculated from the human ranking task (*adequacy* or *fluency*) and from the number of linguistic errors of the corresponding level agree; in a discordant pair, they disagree. The higher the value for τ_B the more similar the linguistic level ranking with the human ranking either in *adequacy* or *fluency*. In all cases, correlations followed a statistically significant trend [10].

Table 5 show the results for 2520 (630 sentences * 4 systems) segments. On the one hand, *adequacy* is clearly correlated to semantics, a little to syntax and nothing to orthography and morphology because these two levels do not interfere in the *understanding* of the translation. On the other hand, *fluency* is correlated with all levels in this order of major to minor importance: semantics, syntax, orthography and morphology. In both cases, *adequacy* and *fluency* are clearly related to the quantity of total errors provided by the system.

4. Conclusions

We proposed an alternative way of human evaluation in machine translation. To the best of our knowledge, our proposal is the first linguistic evaluation which

	Adequacy	Fluency
Orthographic	0	.13
Morphological	0	.11
Semantic	.57	.54
Syntactic	.17	.25
Total errors	.58	.57

Table 5: Correlation at the level of segment between linguistic evaluation and *adequacy* and *fluency*.

.

has been tested in detail providing good inter-annotation agreement, excellent error coverage and informative segment correlation with the standard human evaluation methodology of *adequacy* and *fluency*. In this sense, linguistic guidelines have been shown useful for machine translation evaluation.

This methodology has been proved to achieve a really high inter-annotation agreement (a kappa of 0.75) which should be one of the main goals in machine translation evaluation. The level of agreement achieved is quite surprising specially if we take into account that the evaluation does not use a reference translation.

Moreover, the linguistic guidelines, designed for Spanish-to-Catalan and specific for the target language (Catalan), have shown generalizable for a different source language (English). Annotators could not find one single error that was not specified in the guidelines. Finally, the linguistic classification of errors provides new information which has shown useful to relate linguistic errors from different type of systems. Additionally, we have shown that annotators when evaluating in *adequacy* take into account semantic and syntactic errors and when evaluating in *fluency* take somehow all type of errors into account. Our intention with this correlation analysis was not to reach specially high correlations, but to show how linguistic evaluations are related when studying translation outputs.

In further work, we would like to investigate how these linguistic guidelines work over a crowd-sourcing platform and how this new linguistic information can be used to improve machine translation systems.

5. Acknowledgments

The authors would like to thank the Institute for Infocomm Research for their support and permission to publish this research.

This work has been partially funded by the Seventh Framework Program of the European Commission through the International Outgoing Fellowship Marie Curie Action (IMTraP-2011-29951).

[1] Callison-Burch, C., Koehn, P., Monz, C., Peterson, K., Przybocki, M., Zaidan, O.. Findings of the 2010 joint workshop on statistical ma-

chine translation and metrics for machine translation. In: Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR. Uppsala, Sweden: Association for Computational Linguistics; 2010. p. 17–53.

[2] Coehn, J.. Weighted kappa: Nominal scale agreement and the maximum value of kappa. Educational an Psychological Measurement 1968;(49):835–850.

[3] Costa-jussà, M., Farrús, M., Mariño, J., Fonollosa, J.. Study and comparison of rule-based and statistical catalan-spanish machine translation systems. Accepted in Computing and Informatics journal 2011;.

[4] Farrús, M., Costa-jussà, M., Poch, M., Hernández, A., Mariño, J.. Improving a catalan-spanish statistical translation system using morphosyntactic knowledge. In: 13th Annual meeting of the EAMT: European Association for Machine Translation. Barcelona; 2009. p. 52–57.

[5] Farrús, M., Costa-Jussà, M.R., Popovic, M.. Study and correlation analysis of linguistic, perceptual, and automatic machine translation evaluations. JASIST 2012;63(1):174–184.

[6] Flanagan, M.A.. Error classification for mt evaluation. In: Proc. of the AMTA. Columbia; 1994. p. 65–72.

[7] Kittur, A., Chi, E.H., Suh, B.. Crowdsourcing user studies with mechanical turk. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. New York, NY, USA: ACM; CHI '08; 2008. p. 453–456.

[8] Landis, J.R., Koch, G.G.. The measurement of observer agreement for categorical data. Biometrics 1977;33:157–174.

[9] Mariño, J., Banchs, R., Crego, J., de Gispert, A., Lambert, P., Fonollosa, J., Costa-jussà, M.. N-gram based machine translation. Computational Linguistics 2006;32(4):527–549.

[10] McBride, G.. Anomalies and remedies in non-parametric seasonal trend tests and estimates; 2000. National Institute of Water and Atmospheric Research, Hamilton.

[11] Naskar, S.K., Toral, A., Gaspari, F., Way, A.. A framework for diagnostic evaluation of mt based on linguistic checkpoints. In: Proceedings of the 13th Machine Translation Summit. Xiamen, China; 2011. p. 529–536.

[12] Olive, J.. Global autonomous language exploitation. DARPA/IPTOProposer Information Pamphlet 2005;.

[13] Snover, M., Madnani, N., Dorr, B.J., Schwartz., R.. Ter-plus: paraphrase, semantic, and alignment enhancements to translation edit rate. Machine Translation 2010;23(2-3):117–127.

[14] Snow, R., O'Connor, B., Jurafsky, D., Ng, A.Y.. Cheap and fastbut is it good?: evaluating non-expert annotations for natural language tasks. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2008. p. 254–263.

[15] Su, Q., Pavlov, D., Chow, J.H., Baker, W.C.. Internet-scale collection of human-reviewed data. In: Proceedings of the 16th international conference on World Wide Web. 2007. p. 231–240.

[16] Vilar, D., Xu, J., Fernando-D'Haro, L., Ney, H.. Error analysis of statistical machine translation output. In: Proc. of the LREC. Genoa, Italy; 2006. .

[17] White, J., O'Connell, T., O'Mara, F.. The arpa mt evaluation methodologies: Evolution, lessons, and future approaches. In: Proc. of the 1st Conference of the Association for Machine Translation in the Americas. Columbia; 1994. p. 193–205.

[18] Zhou, M., Wang, B., Liu, S., Li, M., Zhang, D., Zhao, T.. Diagnostic evaluation of machine translation systems using automatically constructed linguistic check-points. In: Proceedings of the 22nd International Conference on Computational Linguistics COLING08. Stroudsburg, PA, USA; volume 1; 2008. p. 1112–1128.