Study and Correlation Analysis of Linguistic, Perceptual, and Automatic Machine Translation Evaluations

Mireia Farrús
Universitat Oberta de Catalunya, Office of Learning Technologies, Barcelona, Spain.
E-mail: mfarrusc@uoc.edu

AQ1 **Marta R. Costa-jussà**
Barcelona Media Innovation Centre, Voice & Language Department, Barcelona, Spain.
E-mail: marta.ruiz@barcelonamedia.org

Maja Popović
Deutsches Forschungszentrum für Künstliche Intelligenz, Language Technology Lab.
E-mail: Maja.Popovic@dfki.de

Carlos Henríquez
Universitat Politècnica de Catalunya, Department of Signal Theory & Communications, Barcelona, Spain.
E-mail: carlos.henriquez@upc.edu

Evaluation of machine translation output is an important task. Various human evaluation techniques as well as automatic metrics have been proposed and investigated in the last decade. However, very few evaluation methods take linguistic aspect into account. In this article, we use an objective evaluation method for machine translation output that classifies all translation errors into one of the five following linguistic levels: orthographic, morphological, lexical, semantic, and syntactic, to analyse its linguistic quality. Linguistic guidelines for the target language are required, and human evaluators use them in to classify the output errors. The experiments are performed on English-to-Catalan and Spanish-to-Catalan translation outputs generated by four different systems: 2 rule-based and 2 statistical. All translations are evaluated using 3 following methods: a standard human perceptual evaluation method, several widely used automatic metrics, and the human linguistic evaluation. Pearson and Spearman correlation coefficients between the linguistic, perceptual, and automatic results are then calculated, showing that the semantic level correlates significantly with both perceptual evaluation and automatic metrics.

Received November 10, 2010; revised July 19, 2011; accepted August 29, 2011

 DOI: 10.1002/asi.21674

Introduction

AQ2

Background

Because machine translation became a popular research field in the 50's, one of the major needs in this area has been to find an appropriate system evaluation procedure to test the quality of the output translations. During the last years, two very different ways of evaluating machine translation systems have appeared within the research community. On the one hand, there are a considerable number of automatic evaluation methods like bilingual evaluation understudy (BLEU; Papieni, Roukos, Ward, & Zhu, 2002), word error rate (WER; McCowan et al., 2004), and translation error rate (TER; Snover, Madnani, Dorr, & Schwartz, 2010). On the other hand, human evaluators have been widely used to analyse the performance of the systems by means of their perception of the translation quality.

Automatic evaluation methods have been providing objective measures to evaluate machine translation systems, where the error rate is measured by comparing the system output against one or several human references. Apart from the above-mentioned methods (BLEU, WER, and TER), other methods include the use of linguistic features (Giménez & Màrquez, 2007; Popovic & Ney, 2009), which make use of linguistic knowledge and correlate with human criteria,

and METEOR (Lavie & Agarwal, 2007), which is becoming quite popular. METEOR is able to produce detailed word-to-word alignments between the system translation and the reference translation, which can help in analysing the errors. The main handicaps of these methods are that manual references cannot cover all possible translations and that some systems can be favored among others depending on the technology used.

Human perceptual evaluations methods are based on a pairwise comparison of systems, where the annotator is asked to choose the best translation. Normally, given a translation output, a source sentence and a reference sentence, the evaluator is asked to score a sentence between 1 and 5 in adequacy and fluency (1 being the lowest score and 5 the highest). Recently, in the DARPA's Global Autonomous Language Exploitation (GALE) program (Olive, 2005), one effective way to evaluate was to ask evaluators to edit the translation by means of human-targeted translation edit rate (HTER), in which the less the number of edits, the better the translation. On the other hand, in Callison-Burch, Koehn, Monz, and Schroeder (2009), the authors proposed to edit the translation output as fluent as possible, which reflects the evaluators' understanding of the sentence. The main handicap of these methods is that although human evaluation does not tend to favour any specific system, human evaluation is always a subjective measure and highly dependent on the inter-annotator agreement.

Motivation and Main Goals

Some proposals regarding evaluation classification schemas can be found in the literature as alternatives to the above-described traditional methods. Vilar, Xu, Fernando-D'Haro, and Ney (2006), for instance, propose a five-category schema that does not rely on any linguistic criterion. The Flanagan classification (Flanagan, 1994) lists a series of errors that are pair language-dependent, and Popovic (2009) presents a framework for automatic error analysis and categorization. The basic idea is to identify erroneous words using algorithms for the calculation of WER and PER. The extracted error details can be used in combination with several types of natural language knowledge, such as base forms, part-of-speech tags, and others. The work focuses on the five error categories from Vilar et al., and the new measures correlate well with the results of human analysis when using the same categorization.

The main objective of the current work is to design an alternative and objective human evaluation method being able to take into account all possible errors for one language and to classify them into linguistic and general categories. Inspired by our previous study in Farrús, Costa-jussà, Mariño, and Fonollosa (2010) and Farrús et al. (2011), an evaluation method based on the assumption that all the errors can be classified into one of the following linguistic levels: orthographic, morphological, lexical, semantic, and syntactic.

Contribution

The current work uses the proposed linguistic evaluation method, aiming at being objective over any translation output and at specifying the type of errors committed by the system. The information extracted by using this evaluation will be useful to the machine translation developers to improve the translation system.

The proposed evaluation can be defined as specific and general at the same time. Specific because it describes all possible errors, and general because all possible errors are classified into general linguistic categories. Also, this proposal intends to be the first attempt to create translation evaluation guidelines for Catalan by identifying all types of errors. Extended guidelines are required for each target language to analyse what type of errors are included in each linguistic level. Human evaluators can use these guidelines to classify and compute the number and the type or errors encountered in the translations to analyse their linguistic quality. Consequently, the current evaluation method can be used to extract information about the nature of the errors committed by a particular system.

Additionally, the human linguistic evaluation is used to find some correlations over linguistic categories and traditional evaluation methods (both automatic and perceptual), so that the linguistic evaluation can provide a further point of view in the evaluation methods. To this end, two language pairs (English into Catalan and Spanish into Catalan) and four different translation systems are considered: two rule-based and two statistical systems. Catalan language is always used as target language because all the evaluators participating in this study were Catalan native.

Structure of the Paper

The structure of this article is as follows. First, the machine translation systems that are later used for experimentation are briefly reviewed. Then we present the proposed linguistic evaluation used in the experiments. In the following section, the experimental setup is described. Next, the results obtained in all the evaluations performed (automatic, perceptual and linguistic) are shown, together with the correlation analysis of the evaluations involved in this study. Finally, the most relevant conclusions are presented.

Machine Translation Paradigms

This section, without aiming at completeness, pretends to briefly introduce the machine translation systems that are used later for experimentation: rule-based, phrase-based, and Ngram-based. We start with some touches of machine translation history and then explain at a high level how these machine translation systems operate. Further details on these systems are out of the scope of this article and the reader can refer to specific papers such as Arnold and Balkan (1995), Dorr (1994), Hutchins (1986), and Lopez (2007).

DOI: 10.1002/asi

The first commercial machine translation systems were the rule-based machine translation (RBMT) systems. The Georgetown-IBM experiment in 1954 can be considered one of the first RBMT systems, and Systran was one of the first companies that developed them. RBMT technology applies a set of linguistic rules in three different phases: analysis, transfer, and generation. Therefore, a rule-based system requires syntactic analysis, semantic analysis, syntactic generation, and semantic generation. One of the main problems in translation is to be able to choose the correct meaning, which involves a classification or disambiguation problem. To improve the accuracy, it is possible to apply a method to disambiguate different meanings of a single word. Machine learning techniques extract automatically the context features that are useful for word disambiguation. Apertium (Forcada, Tyers, & Ramírez, 2009) is a popular example of open-source rule-based system.

On the other hand, given a parallel text at the sentence level, statistical machine translation (SMT) uses probabilistic models to learn translations (Brown, Della Pietra, Della Pietra, & Mercer, 1993). Given a source string (s1J = s1 . . . sj . . . sJ), the goal is to choose the string with the highest probability among all possible target strings (t1I = t1 . . . ti . . . tI). Original word-based translation models have been replaced by phrase-based translation models (Zens, Och, & Ney, 2002; Koehn et al., 2003), which are directly estimated from aligned bilingual corpora by considering relative frequencies. Recent systems implement a general maximum entropy approach in which a log-linear combination of multiple feature functions is used (Och, 2003). This approach leads to maximising a linear combination of feature functions (hm) with their respective weights (λm):

$$\tilde{t} = \arg\max_{t} \left\{ \sum_{m=1}^{M} \lambda_m h_m\,(t, s) \right\} \quad (1)$$

The main model in this combination is the translation model. The objective of this model is, given a target sentence and a source sentence, to assign a probability that s1J generates t1I. While these probabilities can be estimated by thinking about how each individual word is translated, modern SMT is based on the intuition that a better way to compute these probabilities is by considering the behaviour of phrases (sequences of words). The intuition of phrase-based SMT is to use phrases as the fundamental units of translation. Phrases are estimated from multiple segmentations of the aligned bilingual corpora by using relative frequencies. Alternative approaches use a translation model that has been derived from the finite-state perspective (Bangalore & Riccardi, 2000; Casacuberta, 2001; Mariño et al., 2006).

In addition to the translation model, SMT systems use both the language and the lexical models. The former is usually formulated as a probability distribution over strings that attempt to reflect how likely a string occurs inside a language (Chen & Goodman, 1998). SMT systems make use of the same *n*-gram language models, as do speech recognition and other applications. The language model component is monolingual, so that acquiring training data is relatively easy. The lexical models allow the SMT systems to compute another probability to the translation units based on the probability of translating the unit word per word. The probability estimated by lexical models tends to be in some situations less sparse than the probability given directly by the translation model. Many additional feature functions can also be introduced in the SMT framework to improve the translation, like the word or phrase bonus.

Linguistic Evaluation

To carry out the linguistic evaluation for Catalan as target language, an error classification was performed according to the standards of the Institute of Catalan Studies (http://www.iec.cat). Because it is well-known that the same sentence can be translated in many different ways, the following criterion was applied to decide whether a sentence was correct: All the translations achieved can be considered as correct translations if they maintain the meaning of the original sentence and are grammatically correct.

The main errors found in the translation system were classified according to their corresponding linguistic level: orthographic, morphological, lexical, semantic, and syntactic. Based on this classification, some preliminary guidelines were designed using a Spanish-to-Catalan set of 711 sentence (about 16,000 words) extracted from *El País* and *La Vanguardia* newspapers (see Farrús et al., 2011).

For each linguistic level involved in the classification, a list of error subtypes was provided. Most of these errors are language dependent and related to the target language (Catalan in our case). However, some specific errors might depend also on the language pair involved. To cover the pair English-Catalan, the list was extended when a specific error for this language pair was found (the "extra target words," for example). Next, the main specific problems encountered at each linguistic level are described.

Orthographic Errors

The errors related to the orthographic level are as follows:

- **Punctuation marks**: A wrong use, missing punctuation and extra punctuation of exclamation and interrogation marks, full stops, commas, colons, semicolons, dots, etc.
- **Accents**: Accented vowels when not necessary, missing and erroneous accents, e.g., *vosté* instead of *vostè* (meaning *you*).
- **Capital and lower case letters**: 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., *després Encara se sent* instead of *després encara se sent* (meaning *then you still feel*).
- **Joined words**: Two consecutive words erroneously joined, e.g., *ia* instead of *i a*(meaning *and a*).
- **Extra spaces**: Error usually committed due to a nondetokenisation when required or a detokenisation into the wrong direction, e.g., *T' has sentit* instead of *T'has sentit* (meaning *you have felt*).

DOI: 10.1002/asi

- **Apostrophe**: Commonly used in Catalan to elide a sound, in some cases a missing or an extra apostrophe is found when Catalan is the target language, e.g., *el ofec* instead of *l'ofec* (meaning *short of breath*).

Morphological Errors

The errors related to the morphological level are as follows:

- **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*). This problem, found in the Spanish-Catalan pair, is also relevant in the English-Catalan pair, because gender is usually not explicit in English names, adjectives, and articles. 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.
- **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 English and in Spanish (*el dinero*) and plural in Catalan (*els diners*). Like in the gender concordance, this causes a lack of number concordance in articles and adjectives with the consecutive noun.
- **Verbal morphology**: It refers to a verb that is not correctly inflected, a common error in an inflected language at verb level such as Catalan. The most common cases are the translation of an inflected verb into the infinitive form, or the lack of person concordance. This is especially common in the English-Catalan pair, since English is less inflected than Catalan at the verb level, e.g., *does the fever come and go* was translated into *la febre vénen i va* instead of *la febre ve i va* (although the correct expression would be *la febre va i ve*).
- **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.

Lexical Errors

The errors related to the lexical level are as follows:

- **Incorrect words**: No correspondence between the source word and the translated target word. This error is normally found in statistical systems, where the word is translated incorrectly due to training alignment errors, e.g., *a kidney infection* was translated into *això ronyó* instead of *una infecció de ronyó*.
- **Unknown words**: Nontranslated source words, which are left intact in the target side, e.g., *admission and administrative data* was translated into *admission i dades administratives* instead of *admissió i dades administratives.*
- **Missing target words**: Nontranslated source words, which are missing in the target side, e.g., *what brings you here today?* was translated into *vostè porta avui aquí?* Instead of *què el porta a vostè avui aquí?*
- **Extra target words**: Words appearing, for no apparent reason, in the target side, e.g., *administrative data* was translated into *de dades administratives*, where the particle *de* is unnecessary.

Semantic Errors

The errors related to the semantic level include:

- **Polysemy**: The wrong meaning is chosen in the target language when translating a word with multiple meanings (polysemes), e.g., the English word *appointment* was translated by the Catalan word *nomenament* (act of appointing) instead of *cita* (arrangement to meet), which was the correct meaning given the context.
- **Homonymy**: The wrong meaning is chosen in the target language when translating words that share the same spelling—and pronunciation—but have different meanings (homonyms), e.g., the Spanish adverb *solo*, which can also be an adjective, is translated into the Catalan adjective *sol* instead of the corresponding adverb *només* and vice versa. Or the English pronoun *I*, which can also be a number, is translated into the number instead of the corresponding pronoun *jo*.

Syntactic Errors

The errors related to the syntactic level are as follows:

- **Prepositions**: It refers to prepositions not elided in the target language, prepositions not inserted in the target language, or source prepositions maintained in the target language instead of a new correct target preposition, e.g., *in which city* or the equivalent Spanish *en qué ciudad* was translated into *en quina ciutat* instead of *a quina ciutat*. In this case, the Catalan preposition would have been changed with respect English and Spanish languages.
- **Verbal periphrasis**: The use of verbal periphrasis, especially when they involve prepositions that differ in different languages, usually leads to translation errors, as well, e.g., the Spanish verbal periphrasis *tener que* (to have to) is usually translated literally into Catalan as *tener que* instead of the correct periphrasis *haver de*.
- **Clitics**: Include a wrong syntactic function of the pronoun or a wrong clitic-verb combination, e.g., *se ha lesionado* was translated into *es ha lesionat* instead of *s'ha lesionat*.
- **Reordering**: Wrong order of the elements of the sentence. Because of the syntactic differences between Germanic and Romanic languages, this error is more common in the English-Catalan pair, e.g., *Social background I.* Was translated into *I. Context social* instead of *Context social I.*

These main differences between our classification and the works developed by Vilar et al. (2006), Popovic (2009), and Flanagan (1994) are as follows:

(1) A language-dependent error specification is performed. This error specification is ambitious, as it is supposed to include all types of possible errors that can be made in a translation from one language to another. Moreover, it offers more linguistic information about the type of error; e.g., Vilar et al. (2006) use the concept of *incorrect words* that can be related to multiple linguistic levels: lexical, semantic, and morphological.
(2) A linguistic categorization of errors generalizable to other languages is made. This linguistic classification contains five linguistic categories: orthographic, morphological, lexical, semantic, and syntactic. The list of subcategories for Catalan-Spanish is similar to the one

presented in Flanagan (1994); however, our subcategories are included in a five-category schema, which is language independent.

(3) Considering 1 and 2, it can be seen that the linguistic evaluation contains two levels of error classification. The lower level (the detection of all types of possible errors) is language dependent, while the upper level (the linguistic error classification) is language independent.

(4) The current work intends to be a first attempt of translation evaluation guidelines for the Catalan language (being Catalan the target language).

(5) Finally, our linguistic evaluation can be used to investigate which type of error has more influence when evaluating the translation quality. To identify the type of error, a manual evaluation is used instead of an automatic one. Once the errors have been detected and classified, the correlations of our linguistic evaluation with both perceptual and automatic evaluations are calculated, which gives the type of linguistic error that has more importance when looking for translation quality.

Experimental Setup

Next, the four systems used in the current experiments are described. Translations are evaluated using automatic, perceptual, and linguistic criteria, which are studied and compared.

Machine Translation Systems

This section introduces the English-to-Catalan and Spanish-to-Catalan machine translation systems available on the web and used in the current study. They include two RBMT systems, Apertium and Translendium, and two SMT systems, Google Translate and UPC. All the systems are used in their respective date versions of February 1, 2010.

- **Apertium** platform (http://www.apertium.org) 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). It was funded by the *Open-Source Machine Translation for the Languages of Spain* project. Subsequent development has been funded by the UA and by Prompsit Language Engineering.

  Apertium architecture has been released under open-source licenses and distributed free of charge, so that anyone having the necessary computational and linguistic skills will be able to adapt or improve the platform or the language pair data to create a new machine translation or to add a new language pair. Designed according to the UNIX philosophy, the translation is performed in stages by a set of tools operating on a simple text stream. Other tools can be added to the pipeline as required, and the text stream can be modified using standard tools. Because it was initially designed for the translation between related pairs, the system uses a shallow-transfer machine translation technology. In addition, tools for manipulating linguistic data are provided.
- **Translendium** (http://ww.translendium.com) 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 15 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. This engine can be connected to translation memory modules and to a professional lexicon editor. Additionally, it can be accessed through a multiuser task distribution server either from a web client or from a professional single user client. Moreover, the system can be adapted to general, social, technical, and medical documents.
- **Google Translate** (http://translate.google.com) 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 and aligned text comprising examples made by human translators between the languages included.

  Google is constantly working to support more languages and introduce them as soon as the automatic translation meets their standards.
- **UPC** system (http://www.n-ii.org) is developed at the Universitat Politècnica de Catalunya and has been funded by the European Union under the integrated project TC-STAR (Technology and Corpora for Speech to Speech Translation, IST-2002-FP6-506738), and the Spanish Government under the AVIVAVOZ project (Technologies for Speech-to-Speech Translation, TEC2006-13694-C03-01). Based on an N-gram 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 (Farrús et al., 2009).

  English-to-Catalan translation is trained on a Catalan-English parallel corpus where Catalan has been translated from Spanish using the same UPC system for Spanish-to-Catalan. The Spanish-to-Catalan and English-to-Catalan translations have special modules that detect those number and time expressions not included in the training corpus to generate them through linguistic rules. Both systems also have a spell checker to avoid wrong-written—and hence unknown—words as input. Additionally, for Spanish-to-Catalan, other unknown words are solved by including a Spanish–Catalan dictionary as a post-process after the translation.

Corpus

The test corpus is provided within the medicine domain. This medical corpus was kindly provided by the Universal-Doctor project, which focuses on facilitating communication between healthcare providers and patients from various origins (http://www.universaldoctor.com). The medical corpus consists of 630 parallel sentences and only one manual reference for each translation direction was available. Table 1 shows the number of sentences, words and vocabulary used for each language.

AQ2

Evaluation Methods

Automatic metrics. By far, the most widely used metric in the recent literature is the BLEU. It is a quality metric defined in a range between 0 and 1 (or in a percentage between 0

DOI: 10.1002/asi

TABLE 1. Corpus statistics of the trilingual medical English-Spanish-Catalan test set.

	English	Spanish	Catalan
Sentences	630	630	630
Words	4073	3479	3425
Vocabulary	1050	1112	1120

and 100): 0 meaning a bad translation (where the translation does not match the reference in any word), and 1 a supposedly correct translation (according to the available references). AQ3 BLEU most used setting computes lexical matching accumulated precision for n-grams up to length four (Papinieni et al., 2002).

WER (McCowan et al., 2004) is a standard speech recognition evaluation metric. A general difficulty of measuring performance lies in the fact that the translated word sequence can have a different length from the reference word sequence (supposedly the correct one). The WER is derived from the Levenshtein distance, working at the word level.

On the other hand, TER is an error metric for machine translation that measures the number of edits required to change a system output into one of the references (Snover et al., 2010). This measure is similar to WER, but with additional shift costs.

Note that, although METEOR is also a widely-used automatic evaluation metric, it requires specific training and is not yet done for the Catalan language.

AQ4 Although automatic evaluation is a must in MT to train the systems, some of the main problems are that the measure depends on the references quality and the measure does not AQ4 behave objectively among different types of MT translation systems (i.e., BLEU favors SMT systems rather than rule-based ones (Callison-Burch, Osborne, & Kohen, 2006). In addition, given that a source sentence might have multiple correct target sentences, it becomes difficult to compose a test set that covers all of them.

Human perceptual evaluation. The comparison between the translation system outputs was performed by 12 different evaluators. All of them were bilingual in Catalan and Spanish and fluent in English; therefore, no translation reference was shown to them to avoid any bias in their evaluation.

The evaluators performed the following comparison. Each judge was asked to make a system-to-system (pairwise) comparison (see Figure 1). Each annotator evaluated 2,835 randomly extracted translation pairs, and assessed in each case whether the translation of one of the systems was better than the other one, or whether both outputs were equivalent. Figure 1 shows an example of the screenshot shown to the annotator. Each judge did such evaluation for three system pairs, so that a total number of 34,020 ($630 \cdot 18 \cdot 3$) judgements was collected, i.e., three different evaluators for each pair of systems. The inter-annotator agreement in this case equalled 80% and 85% for English-to-Catalan and Spanish-to-Catalan directions, respectively. Therefore, if we measure

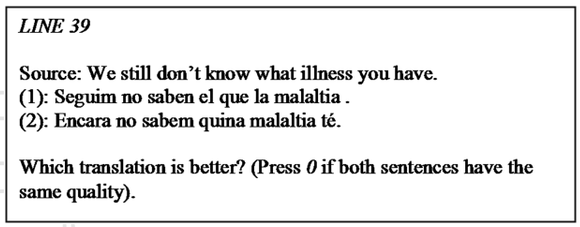

FIG. 1. Screenshot of the human evaluation when comparing two different systems.

the pairwise agreement among evaluators using the Kappa coefficient (K) defined as:

$$\mathrm{K} = \frac{P(A) - P(E)}{1 - P(E)} \tag{2}$$

where $P(A)$ is the proportion of times that the evaluators agree (in our case 0.70 and 0.78) and $P(E)$ the proportion of times that they would agree by chance (in our case 0.5), K equals 0.60 and 0.71, respectively. The interpretation of Kappa is quite variable, but according to Landis and Koch (1977), a value between 0.6 and 0.8 is considered a good inter-annotation agreement.

One of the reasons for which such reasonable Kappa coefficients are obtained might be that the evaluators are bilingual in Spanish and Catalan, and fluent in English. Additionally, the pairwise method is quite clear for evaluators. In fact, the objective of using this human evaluation method was to obtain a high inter-annotation agreement.

Human linguistic evaluation. The linguistic evaluation was performed by five evaluators, who were bilingual in Catalan and Spanish and fluent in English. The errors are reported according to the following linguistic levels, as described above: orthographic, morphological, lexical, semantic, and syntactic. Although the guidelines were designed on a different set from the test set, most of the errors found by the evaluators were reported in these guidelines. In some minor cases, the new errors were added to the list of subtype errors.

The inter-annotation agreement was evaluated using the weighted kappa coefficient (Cohen, 1968), using a linear unitary distance between errors. The weighted kappa equalled 0.64, which is a good value according to Landis and Koch (1977). Note that this inter-annotation agreement is much higher than the ones obtained by popular human evaluation methods such as the ones in Callison-Burch et al. (2009). AQ5

Correlation methods. With the purpose of finding a relationship between linguistic, perceptual, and automatic evaluations, two different correlation methods were considered: Pearson linear correlation and Spearman rank correlation (Edwards, 1976; Spearman, 1904).

The Pearson correlation measures the linear dependency between two variables X and Y, giving values between -1 and 1, inclusive. A value equalling 1 implies that the variables are directly proportional, i.e., Y increases as X does.

On the other hand, a value equalling −1 means that the variables are inversely proportional, i.e., Y decreases when X increases. A value equalling 0 means that $X$and Y are linearly independent. The Pearson correlation is defined as the covariance of both variables divided by the product of their standard deviation:

$$\rho_{X,Y} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_y}. \quad (3)$$

Substituting estimates of the covariances and variances based on a sample, the sample correlation coefficient is obtained:

$$r_{X,Y} = \frac{\sum_{i=1}^{N}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{N}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{N}(Y_i - \overline{Y})^2}} \quad (4)$$

The Spearman rank correlation also estimates how well two variables are related, but the relation should not be necessarily linear. It can be defined as the Pearson correlation between the ranked variables. The difference between both correlations is that the Spearman correlation shows whether the variables are monotonically related, even if their relationship is not linear. The Spearman correlation can be obtained by using Equation 2, but replacing X_i and Y_i by their ranked values x_i and y_i, and $\overline{X}$ and $\overline{Y}$ by the sample mean of x and y. Because the Spearman correlation is defined by the same formula than the Pearson correlation, it also gives values between −1 and 1, inclusive, with an analogous interpretation: positive values mean that the variable Y tends to increase when the variable X increases and negative values stand for a decrease in Y when X increases.

Evaluation Results

This section shows the evaluation results obtained in the three different evaluations performed: automatic, perceptual, and linguistic, as well as the correlation analysis between them.

Automatic evaluation results

The scores obtained by the systems using the automatic evaluation are shown in Tables 2 and 3. For the English-to-Catalan translation (Table 2) the best performing systems are Google and Translendium, which obtained a BLEU score equalling 21.41 and 16.99, respectively. For the Spanish-to-Catalan task (Table 3), the best systems are Translendium and UPC, which obtained BLEU scores equalling 60.92 and 60.69, respectively. In both tasks the worst score was obtained in the Apertium system, which achieved 10.66 BLEU points in the English-to-Catalan task and 55.21 BLEU points in the Spanish-to-Catalan one.

The Spanish-to-Catalan performance (Table 3) is better than the English-to-Spanish ones (Table 2). This might be explained by the fact that the first task is easier than the second one. Spanish and Catalan languages belong to the same family of languages (Romanic) and they are quite similar in the most general terms. English and Catalan, however, do not belong to the same family of languages (English is a Germanic language), and so they are more different in all their linguistic levels and usually report lower BLEU values in the translation tasks.

TABLE 2. Automatic evaluation results for English-to-Catalan translation outputs.

English-to-Catalan (%)	BLEU	TER	WER
Apertium	10.66	73.98	74.51
Google	21.41	62.42	62.91
Translendium	16.99	63.91	64.59
UPC	12.59	68.78	69.07

TABLE 3. Automatic evaluation results for Spanish-to-Catalan translation outputs.

Spanish-to-Catalan (%)	BLEU	TER	WER
Apertium	55.21	28.06	29.91
Google	60.22	26.82	27.35
Translendium	60.92	25.94	26.42
UPC	60.69	25.82	26.33

Note. BLEU = bilingual evaluation understudy; TER = translation error rate; WER = word error rate.

Human Perceptual Evaluation Results

As mentioned before, the human perceptual evaluation comprises a pairwise comparison of the systems. The systems had a similar performance from 10% to 30% of the cases in the English-to-Catalan translation, depending on the pair of systems evaluated, and from 70% to 75% of the cases in the Spanish-to-Catalan translation. The results are shown in Tables 4 and 5. It can be seen that the results are more polarized in the English-to-Catalan case than in the Spanish-to-Catalan case.

Given that Spanish and Catalan are family-related languages, the translation task is easier and there are a large number of cases in which the outputs coincide.

Linguistic Evaluation Results

The number and type of linguistic errors obtained in the linguistic evaluation English-to-Catalan and Spanish-to-Catalan are shown in Tables 6 and 7, respectively. The results show that in both English-to-Catalan and Spanish-to-Catalan translations, the least frequent errors were the orthographic errors (101 and 18, respectively), while the most frequent errors were the semantic errors in the English-to-Catalan translation (1020), and the syntactic errors in the Spanish-to-Catalan translation (187).

Considering the core technology of the translation systems, it can be seen in Table 6 that the SMT systems (Google and UPC) have a percentage of semantic errors below 30%

DOI: 10.1002/asi

TABLE 4. Human evaluation results for English-to-Catalan translation outputs: pair wise comparison.

English-to-Catalan pairwise comparison (%)				
A	*B*	*A* better than *B*	*B* better than *A*	*A* equal to *B*
Apertium	Google	14.4	64.4	21.2
Apertium	Translendium	10.8	61.2	28.0
Apertium	UPC	34.0	34.0	32.0
Google	Translendium	45.6	37.6	16.8
Google	UPC	62.8	18.8	18.4
Translendium	UPC	63.6	17.2	19.2

TABLE 5. Human evaluation results for Spanish-to-Catalan translation outputs: pair wise comparison.

Spanish-to-Catalan pairwise comparison (%)				
A	*B*	*A* better than *B*	*B* better than *A*	*A* equal to *B*
Apertium	Google	8.8	21.2	70.0
Apertium	Translendium	8.8	18.2	73.0
Apertium	UPC	6.8	22.2	71.0
Google	Translendium	16.0	15.2	68.8
Google	UPC	10.0	15.8	74.2
Translendium	UPC	17.8	18.6	63.6

TABLE 6. Linguistic evaluation results for English-to-Catalan translation outputs: number and type of linguistic errors.

English-to-Catalan	sent. with errors	total errors	ort.	mor.	lex.	sem.	Syn.
Apertium	464	731	10	79	121	342	179
Google	305	492	27	72	87	145	161
Translendium	324	478	31	30	65	228	124
UPC	519	1168	33	139	410	305	281

AQ16

TABLE 7. Linguistic evaluation results for Spanish-to-Catalan translation outputs: number and type of linguistic errors.

Spanish-to-Catalan	sent. with errors	total errors	ort.	mor.	lex.	sem.	Syn.
Apertium	112	123	0	3	23	41	56
Google	107	120	5	20	18	17	60
Translendium	73	82	9	0	15	26	32
UPC	84	97	4	16	22	16	39

AQ16

(i.e., 145/492 for Google and 305/1168 for UPC). The rule-based systems (Apertium and Translendium) contain a much higher percentage of semantic errors, over 46% (i.e., 342/731 for Apertium and 228/472 for Translendium). The same happens in the case of Spanish-to-Catalan, where the SMT systems have a lower percentage of semantic errors (less than 17) and the rule-based systems have a higher percentage (over 31).

In the English-to-Catalan direction of translation, the UPC system is not very good in terms of quantity of errors. However, human evaluators in perceptual evaluation do not take these errors as the most important ones. That is why human evaluation and linguistic evaluation do not agree

On the other hand, the SMT systems have a higher relative percentage of morphological errors: in English-to-Catalan, Google has 14.6% and UPC 11.3%, whereas Apertium has 10.8% and Translendium has 6.4%; in Spanish-to-Catalan, both Google and UPC have 16.7%, Apertium has 2.5%, and Translendium has 0%. It might seem surprising to get 0 errors. However, notice that 0 errors appear in the Spanish-to-Catalan task, which is an easier task given the similarity between Spanish and Catalan.

To sum up, the least frequent errors committed by all the systems in both translation directions are the orthographic ones. Likewise, the most frequent errors are found in the semantic and syntactic levels.

TABLE 8. English-to-Catalan Pearson (L) and Spearman (R) correlations.

	Hum.		Ort.		Mor.		Lex.		Sem.		Syn.		all		BLEU		TER		WER	
	L	R	L	R	L	R	L	R	L	R	L	R	L	R	L	R	L	R	L	R
Hum.	1	1																		
Ort.	0.47	−1	1	1																
Mor.	−0.68	0.6	0.12	−0.6	1	1														
Lex.	−0.62	0.6	0.37	−0.6	0.93	1	1	1												
Sem.	−0.94	1	−0.47	−1	0.43	0.6	0.47	0.6	1	1										
Syn.	−0.69	0.8	0.21	−0.8	0.99	0.8	0.98	1	0.49	0.8	1	1								
all	−0.79	0.6	0.15	−0.6	0.93	1	0.97	1	0.65	0.6	0.97	0.8	1	1						
BLEU	0.95	1	0.47	−1	−0.45	0.6	−0.48	0.6	−0.99	1	−0.50	0.8	−0.66	0.6	1	1				
TER	−0.94	1	−0.72	−1	0.41	0.6	0.31	0.6	0.94	1	0.40	0.8	0.53	0.6	−0.94	1	1	1		
WER	−0.93	1	−0.73	−1	0.38	0.6	0.29	0.6	0.94	1	0.38	0.8	0.51	0.6	−0.94	1	0.99	−1	1	1

Note. BLEU = bilingual evaluation understudy; TER = translation error rate; WER = word error rate.

TABLE 9. Spanish-to-Catalan Pearson (L) and Spearman (R) correlations.

	Hum.		Ort.		Mor.		Lex.		Sem.		Syn.		all		BLEU		TER		WER	
	L	R	L	R	L	R	L	R	L	R	L	R	L	R	L	R	L	R	L	R
Hum.	1	1																		
Ort.	0.74	−0.4	1	1																
Mor.	0.42	0	−0.11	−0.2	1	1														
Lex.	−0.47	0.4	−0.93	−1	0.19	0.2	1	1												
Sem.	−0.91	1	−0.52	−0.2	−0.76	−0.6	0.32	0.2	1	1										
Syn.	−0.60	0.6	−0.66	−0.4	0.43	0.8	0.40	0.4	0.21	0.6	1	1								
all	−0.74	0.8	−0.80	−0.8	0.28	0.4	0.55	0.8	0.39	0.8	0.82	0.8	1	1						
BLEU	−0.97	1	0.85	−0.8	0.38	0.4	−0.65	0.8	−0.89	0.8	−0.55	0.8	−0.72	1	1	1				
TER	−0.97	1	−0.79	−0.4	−0.2	0	0.51	0.4	0.78	1	0.77	0.6	0.88	0.8	−0.94	1	1	1		
WER	−0.99	1	−0.82	−0.4	−0.3	0	0.57	0.4	0.85	1	0.68	0.6	0.82	0.8	−0.98	1	0.99	1	1	1

Note. BLEU = bilingual evaluation understudy; TER = translation error rate; WER = word error rate.

Correlation Analysis

In this section, a correlation analysis between the linguistic evaluation, the standard automatic measures, and the human pairwise comparison of the systems is made. Once we obtained the automatic, perceptual, and linguistics results, we grouped all these metrics—a total of 10—and defined a vector for each of them. Each vector contains four different values, one for each MT system considered. Then, two correlation matrices were computed for these sets of vectors, one for Pearson correlation and the other for Spearman.

Tables 8 and 9 show the different correlation matrices for both translation tasks. Considering that two variables are related if they obtain a p-value (p) equal or smaller than 0.05, the following results can be seen.

In the English-to-Catalan task (Table 8):

- Pearson correlation relates syntactic, morphological, and lexical errors.
  - Spearman correlation relates lexical and all the linguistic errors.
  - Pearson correlation shows a relationship between perceptual, BLEU, and semantic errors.
  - Spearman correlation not only shows the same, but also adds the rest of automatic measures and orthographic errors into the relationship.
  - Pearson and Spearman correlations related BLEU, TER, and WER.
- In the Spanish-to-Catalan task (Table 9):
  - Spearman correlation relates orthographic and lexical errors.
  - Spearman correlation matrix also relates the perceptual evaluation, automatic measures, and the semantic errors and almost does the Pearson correlation.
  - Pearson relates automatic measures and perceptual evaluation.
  - Pearson and Spearman correlations relate BLEU, TER, and WER.

Both correlation methods agree in both language pairs in the fact that semantic errors, perceptual evaluation, and automatic metrics are related. Additionally, the correlation between morphological, lexical, and syntactic errors is different for both tasks. This can be explained by the fact that Spanish and Catalan are family-related languages, whereas English and Catalan are not. The similarity between

DOI: 10.1002/asi

languages reduces the differences in evaluation and there are therefore less errors, and significance in correlation is more difficult to achieve.

Conclusions

This article presents a study of three evaluation methods applied on four rule-based and statistical machine translation systems on the English-to-Catalan and Spanish-to-Catalan translation directions. The first two methods are based on the traditional automatic and perceptual evaluation measures, while the third one is a new proposed linguistic evaluation method (specific for the target language) based on the errors committed by the systems considering different linguistic levels: orthographic, morphological, lexical, semantic, and syntactic. This linguistic evaluation presents the following advantages compared with existing evaluations: (a) it is objective because all types of errors are specified into predefined guidelines with a good inter-annotator agreement (weighted kappa of 0.65); and (b) it gives linguistic information about the errors committed by the system. Although the linguistic evaluation requires particular guidelines for each target language, it is generalizable to other languages, because all types of errors are classified into general linguistic categories.

The experiments in this article report that SMT systems tend to commit less relative semantic errors than RBMT systems, whereas RBMT tend to commit less relative morphological errors than SMT systems. The linguistic evaluation shows that the least frequent errors committed by all the systems on both translation directions are the orthographic ones. Likewise, the most frequent errors are found on the semantic and syntactic levels.

Furthermore, a correlation analysis has been carried out to see whether and in which degree the linguistic evaluation correlates with standard automatic and perceptual evaluation methods. The analysis used two different correlations: the Pearson lineal correlation and the Spearman rank correlation. Although the results obtained in both translation directions are not exactly the same, they share some coincidences from which the following conclusions can be stated: The semantic level, the perceptual evaluation, and the automatic evaluation measures tend to be correlated. However, perceptual and automatic evaluations do not seem to be correlated with other linguistic levels than the semantic one. This analysis also showed a high correlation between the morphological and lexical levels in the English-to-Catalan translation, which seems to be that errors in morphology lead to errors in the lexis and vice versa. Likewise, in the Spanish-to-Catalan translation, a high correlation was found between the orthographic and the lexical levels, so that errors in orthography lead to errors in lexis and vice versa.

Further research will test the linguistic evaluation in a more challenging annotation environment like the Amazon Mechanical Turk, where evaluators are not linguistic experts. Additionally, our linguistic evaluation can be useful to see if alternative available automatic measures evaluation other linguistic levels rather than the semantic one. Furthermore, this work tried to present a preliminary overview of the correlation between automatic and linguistic analysis. However, it would be useful to corroborate the results by using a bigger test-set and other pairs of languages as future work.

Acknowledgments

This work has been partially funded by the Spanish Department of Science and Innovation through the *Juan de la Cierva* fellowship program and the BUCEADOR project (TEC2009-14094-C04-01). The authors also want to thank the Barcelona Media Innovation Centre for its support and permission to publish this research.

References

Arnold, D., & Balkan, L. (1995). Machine translation: An introductory guide. Computational Linguistics 21(4), 577–578.

Bangalore, S., & Riccardi, G. (2000). Stochastic finite-state models for spoken language machine translation. Proceedings of the Workshop on Embedded Machine Translation Systems (pp. 52–59).

Brown, P., Della Pietra, S., Della Pietra, V., & Mercer, R. (1993). The mathematics of statistical machine translation. Computational Linguistics, 19(2), 263–311.

Callison-Burch, C., Kohen, P., Monz, C., & Schroeder, J. (2009). Findings of the 2009 Workshop on Statistical Machine Translation. Proceedings of the Fourth Workshop on Statistical Machine Translation (pp. 1–28). Stroudsburg, PA: Association for Computational Linguistics.

Callison-Burch, C., Osborne, M., & Kohen, P. (2006). Re-evaluating the role of BLEU in machine translation research. In Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (pp. 249–256). Stroudsburg, PA: Association for Computational Linguistics.

Casacuberta, F. (2001). Finite-state transducers for speech-input translation. In Proceedings of the Automatic Speech Recognition and Understanding Workshop, ASRU (pp. 375–380). Washington, DC: IEEE Press.

Chen, S.F., & Goodman, J.T. (1998). An empirical study of smoothing techniques for language modelling (Tech. Rep. No.). Cambridge, MA: Harvard University.

Cohen, J. (1968). Weighed kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin, 70(4), 23–220.

Dorr, B.J. (1994). Machine translation: A view from the lexicon. Computational Linguistics, 20(4), 670–676.

Edwards, A.L. (1976). The correlation coefficient. An introduction to linear regression and correlation (chap. 4, pp. 36–46). San Francisco: W.H. Freeman.

Farrús, M., Costa-jussà, M., Mariño, J., & Fonollosa, J.A.R. (2010). Linguistic-based evaluation criteria to identify statistical machine translation errors. Proceedings of the 14th Annual Meeting of the European Association for Machine Translation (pp. 167–173). Saint-Raphaël, France.

Farrús, M., Costa-jussà, M.R., Poch, M., Hernández, A., Mariño, J.B., & Fonollosa, J.A.R. (2009). Improving a Catalan-Spanish statistical translation system using morphosyntactic knowledge. In Proceedings of the 13th Annual Meeting of the European Association for Machine Translation (pp. 52–57). Berlin, Germany: Springer.

Farrús, M., Costa-jussà, M., Mariño, J., Poch, M., Hernández, A., Henríquez, C., & Fonollosa, J.A.R. (2011). Overcoming statistical machine translation limitations: error analysis and proposed solutions for the Catalan-Spanish language pair. Language Resources and Evaluation (forthcoming).

Flanagan, M.A. (1994). Error classification for MT evaluation. Proceedings of the Association for Machine Translation in the Americas (pp. 65–72). Maryland.

Forcada, M.L., Tyers, F.M., & Ramírez, G. (2009). The apertium machine translation platform: Five years on. First International Workshop on Free/Open-Source Rule-Based Machine Translation. Alacant, Spain.

Giménez, J., & Màrquez, L. (2007). Linguistic features for automatic evaluation of heterogeneous MT systems. Proceedings of the Second Workshop on Statistical Machine Translation (pp. 256–264). Stroudsburg, PA: Association for Computational Linguistics.

Hutchins, W.J. (1986). Machine translation: Past, present, future. Chichester, UK: Ellis Horwood.

Koehn, P., Och, F., & Marcu, D. (2003). Statistical phrase-based translation. Proceedings of the Human Language Technology Conference (HLT-NAACL) (pp. 127–133). Stroudsburg, PA: Association for Computational Linguistics.

Landis, J.R., & Koch, G.G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33, 159–174.

Lavie, A., & Agarwal, A. (2007). METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. Proceedings of the Second Workshop on Statistical Machine Translation (pp. 65–72). Stroudsburg, PA: Association for Computational Linguistics.

Lopez, A. (2007). A survey of statistical machine translation. Storming Media.

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

McCowan, I., Moore, D., Dines, J., Gatica-Pérez, D., Flynn, M., Wellner, P., & Bourlard, H. (2004). On the use of information retrieval measures for speech recognition evaluation. IDIAP-RR 04-73. Martigny, Switzerland: IDIAP.

Och, F. (2003). Minimum error rate training in statistical machine translation. Proceedings of the 41th Annual Meeting of the Association for Computational Linguistics (pp. 160–167). Stroudsburg, PA: Association for Computational Linguistics.

Olive, J. (2005). Global autonomous language exploitation (GALE), DARPA/IPTO Proposer Information Pamphlet.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (pp. 311–318). Stroudsburg, PA: Association for Computational Linguistics.

Popović, M. (2009). Machine translation: Statistical approach with additional linguistic knowledge. Unpublished doctoral dissertation, RWTH Aachen University, Aachen, Germany.

Popović, M., & Ney, H. (2009). Syntax-oriented evaluation measures for machine translation output. Proceedings of the Fourth Workshop on Statistical Machine Translation (pp. 29–32). Stroudsburg, PA: Association for Computational Linguistics.

Snover, M., Madnani, N., Dorr, J., & Schwartz, R. (2010). TER-plus-paraphrase, semantic, and alignment enhancements to translation edit rate. Machine Translation, 23(2–3), 117–127. Berlin, Germany: Springer.

Spearman, C. (1904). The proof and measurement of association between two things. American Journal of Psychology, 15, 72–101.

Vilar, D., Xu, J., Fernando-D'Haro, L., & Ney, H. (2006). Error analysis of statistical machine translation output. Proceedings of the International Conference on Language Resources and Evaluation (pp. 697–702). Genoa, Italy.

Zens, R., Och, F., & Ney, H. (2002). Phrase-based statistical machine translation. In Jarke, M., Koehler, J., & Lakemeyer (Eds.), KI-2002: Advances in Artificial Intelligence (Vol. 2479, pp. 18–32). Aachen, Germany: Springer Verlag.

Author Queries

AQ1: Please provide complete mailing addresses for each author.
AQ2: Please provide an introductory sentence for this section.
AQ3: This phrase is unclear–please revise it for greater clarity.
AQ4: Does MT stand for machine translation? If so, spell it out instead, as the acronym has not been used in the article.
AQ5: Please revise for greater clarity: “The inter-annotation agreement was evaluated using the weighted kappa coefficient (Cohen, 1968) and a linear unitary…” or “The inter-annotation agreement was evaluated by using both the weighted kappa coefficient (Cohen, 1968) and a linear unitary…”
AQ6: Please add the comparison; perhaps: “more different than Spanish and Catalan” if that’s what you mean.
AQ7: Please revise into complete sentences, such as: “The linguistic evaluation results for Spanish-to-Catalan task are as follows:” and “The Spanish-to-Catalan correlations are as follows:”
AQ8: Please clarify: should this be “as does the Pearson correlation?" (also, note that ‘Pearson’s’ has been changed to ‘Pearson’ throughout the article, per APA style)
AQ9: The meaning is unclear–please revise this passage for greater clarity.
AQ10: For every proceedings, please include the title of the article, editors, title of the conference (spell it out in full, followed by the abbreviated title and year in parentheses), page numbers, publisher location and publisher. Readers might also appreciate a link to the work.
AQ11: Please provide the tech report number after ‘No.’ and the department after the university.
AQ12: Per APA style, delete ‘forthcoming’ and replace the year with ether in press or “Manuscript submitted for publication” both here and in the in-text citations.
AQ13: Please provide the location (city/state) of this publisher (or a link).
AQ14: Please provide complete reference info for your readers–a link, or publisher & publisher location.
AQ15: The first author’s name is spelled differently in the citations–please make the correction where appropriate.
AQ16: Add the yellow-highlighted abbreviations to the footnote sections of tables 6-9.
AQ17: Is this a technical report?