Noname manuscript No.
(will be inserted by the editor)

A Linguistically Motivated Taxonomy for Machine Translation Error Analysis

Ângela Costa · Wang Ling · Tiago Luís · Rui Correia · Luísa Coheur

Received: date / Accepted: date

Abstract A detailed error analysis is a fundamental step in every natural language processing task, as to be able to diagnosis what went wrong will provide cues to decide which are the research directions to be followed. In this paper we focus on error analysis in Machine Translation. We deeply extend previous error taxonomies so that translation errors associated with Romance languages specificities can be accommodated. Also, based on the proposed taxonomy, we carry out an extensive analysis of the errors generated by four different systems: two mainstream online translation systems Google Translate (Statistical) and Systran (Hybrid Machine Translation) and two in-house Machine Translation systems, in three scenarios representing different challenges in the translation from English to European Portuguese. Additionally, we comment on how distinct error types differently impact translation quality.

Keywords Machine Translation · Error Taxonomy · Error Analysis · Romance Languages

1 Introduction

Error Analysis is the process of determining the incidence, nature causes and consequences of unsuccessful language (James, 1998). This linguistic discipline has been applied to many research fields, such as *Foreign Language Acquisition* and *Second Language Learning and Teaching* (Corder, 1967), since errors contain valuable information on the strategies that people use to acquire a language (H. Dulay and Krashen, 1982) and, at the same time, allow to identify points that need further work. In fact, according to Richards (1974), "At the level of pragmatic classroom experience, error analysis will continue to provide one means by which

INESC-ID
Rua Alves Redol, 9
1000-029 Lisbon
Tel.: +351-21-3100300
Fax: +351-21-3100235
E-mail: {angela, wlin, tiago.luis, rui.correia, luisa.coheur}@l2f.inesc-id.pt

the teacher assesses learning and teaching and determines priorities for future effort". More recently, error analysis has also become a focus of research in the Machine Translation (MT) area, where some works are dedicated to the design of taxonomies (Llitjós et al, 2005; Vilar et al, 2006; Bojar, 2011) and others target errors' identification (Popović and Ney, 2006). In this paper, **we present a linguistically motivated taxonomy for translation errors that extends previous ones.** Contrary to other approaches, our proposal:

- clusters different types of errors in the main areas of linguistics, allowing to precise the information level needed to identify the errors and easing a possible extension process;
- allows to classify errors that occur in Romance languages and not only English (usually ignored in previous taxonomies);
- allows to take into consideration language's variations.

Moreover, based on this taxonomy **we perform a detailed linguistic analysis of the errors produced in the translation of English (EN) into European Portuguese (EP) texts by two mainstream online translation systems, Google Translate and Systran, and two in-house MT systems, both based on Moses technology**.

Google Translate[1], provided by Google, is the best known translation engine, allowing to perform translations of texts in many languages; Systran[2] is a free online hybrid machine translation engine that combines rule-based and statistical machine translation. Moses[3] is a publicly available statistical machine translation system, intensively used by MT researchers all over the world. It allows to automatically train translation models for any language pair, as long as a collection of parallel corpus is available, such as the Europarl (Koehn, 2005).

Due to the fact that MT statistical systems are highly dependent of the training data and, thus, behave differently in distinct domains, we have chosen parallel corpora with different characteristics. Therefore, we perform our experiments in a corpus that contains the one described in (Costa et al, 2014), and is composed of:

- speech transcriptions (and respective translations of the subtitles into EP) of TED-talks[4];
- touristic texts from the bilingual UP magazine[5];
- TREC evaluation questions (Li and Roth, 2002), translated into EP in a previous work by Costa et al (2012)[6].

In this way, we are able to study and cover errors resulting, respectively, from speech translations, from translations within a restricted domain and also from translations over specific constructions, which is the case of questions. Moreover, the EP translations of the corpora were used to automatically evaluate the translation performed by all systems.

This paper is organised as follows: in Section 2 we present related work, in Section 3 we detail the error taxonomy, and, in Section 4, we describe the corpora,

[1] http://translate.google.com
[2] http://www.systranet.com/translate
[3] http://www.statmt.org/moses
[4] http://www.ted.com/
[5] http://upmagazine-tap.com/
[6] http://metanet4u.l2f.inesc-id.pt/

the tools used in our experiments and the annotation process. In Section 5 we analyze the errors resulting from the translations and in Section 6 we discuss error gravity. Finally, in Section 7, we highlight the main conclusions and point to future work.

2 Related Work

Several studies have been developed with the goal of classifying translation errors in MT. In addition, several works focus on the identification of machine or human translation errors. Some researches target semi- or fully- automatic error analysis methods, others manually analyze these errors. In this section we survey the most significant work on these subjects.

Starting with the problem of errors' classification, different taxonomies have been suggested. One of the most referred taxonomies in MT is the hierarchical classification proposed by Vilar et al (2006). They extend the work of Llitjós et al (2005), and split errors into five classes: "Missing Words", "Word Order", "Incorrect Words", "Unknown Words" and "Punctuation Errors". A "Missing Words" error is produced when some words in the generated sentence are missing. "Word Order" errors concern the word order of the generated sentence. This problem is solved by moving words or blocks of words within the sentence. "Incorrect Words" are found when the system is unable to find the correct translation of a given word. "Unknown Words" are "translated" simply by copying the input word to the generated sentence, without further processing. Finally, "Punctuation Errors" represent only minor disturbances, but are also considered in this taxonomy.

Inspired by the work of Vilar et al (2006), Bojar (2011) used a similar classification that divides errors into four types: "Bad Punctuation", "Missing Word", "Word Order" and "Incorrect Words". Basically he uses Vilar's taxonomy, but eliminates the "Unknown Words" category.

The classification of errors done by Elliott et al (2004) was progressively developed during the analysis and manual annotation of approximately 20.000 words of MT output, translated from French into English by four systems (Systran[7], Reverso Promt[8], Comprendium[9] and SDL's online Free Translation[10]). This taxonomy is slightly different, as the annotations were made according to items that a post-editor would need to amend if he/she was revising the texts to publishable quality. Error types were divided according to parts-of-speech and then sub-divided as "Inappropriate", "Untranslated", "Incorrect", "Unnecessary' and "Omitted".

At this point it is important to say that all taxonomies are influenced by the idiosyncrasies of the languages with which they are working. For instance, the work carried out by Vilar et al (2006) concerns experiments with the language pair English-Chinese and, thus, takes into consideration error types that are not relevant for European languages. For instance, working particularly with Chinese they felt the need to introduce new types of reordering errors, as the position of the modifier changes according to the sentence construction (declaratives, interrogatives, subordinates/infinitives sentences). In our particular case, as we are

[7] http://www.systranet.com/translate
[8] http://reverso.softissimo.com/en/reverso-promt-pro
[9] http://amedida.ibit.org/comprendium.php
[10] http://www.freetranslation.com

working with translations from EN into EP, our main concern was to develop a taxonomy that captures all idiosyncrasies of Portuguese but that also works for Romance languages.

Although the purpose of this work is to classify machine translation errors, for the creation of our error taxonomy, we think it is also important to consider the error classification studies for human errors. In what concerns human translation errors, H. Dulay and Krashen (1982) suggest two major descriptive error taxonomies: the Linguistic Category Classification (LCC) and the Surface Structure Taxonomy (SST). LCC is based on linguistic categories (general ones, such as morphology, lexis, and grammar and more specific ones, such as auxiliaries, passives, and prepositions). SST focuses on the way surface structures have been altered by learners (e.g., omission, addition, misformation, and misordering). These two approaches are presented as alternative taxonomies. However, according to James (1998) there is a great benefit to combining them into a single bidimensional taxonomy.

Also concerning human errors, but this time errors produced by humans in a translation task, we should mention the Multilingual eLearning in Language Engineering project[11] (MeLLANGE). They produced the MeLLANGE Learner Translator Corpus that includes work done by trainees, which was subsequently annotated for errors according to a customised error typology. We still have not done any experiments with human translation errors, that we have planned for the future, but we also had this type of errors in mind, when creating our taxonomy.

Considering the identification of MT errors, several automatic measures are proposed in the literature. Among these, two of the most widely used scores in Statistical MT are the Bilingual Evaluation Understudy (BLEU) (Papineni et al, 2002), and METEOR (Denkowski and Lavie, 2014). BLEU scores are calculated by comparing translated segments with reference translations. Those scores are then averaged over the whole corpus to reach an estimate of the translation's overall quality. BLEU simply calculates N-gram precision adding equal weight to each one (an usual critic to BLEU is that it does not take into account intelligibility or grammatical correctness). Finally, METEOR (Denkowski and Lavie, 2014) is an automatic metric for machine translation evaluation that is based on a generalized concept of unigram matching between the machine-produced translation and human-produced reference translations. It also uses other linguistic resources such as paraphrases and generally obtains better results that is why we have chosen to use it in our automatic evaluation. However, even though automatic evaluation methods are very much desired as they are quicker and less expensive than a manual evaluation, human judgements of translation performance are still more accurate. We should also add that the interpretation of these measures is not easy and they do not permit a clear identification of the engines' problems. For instance, a BLEU score of 0.20 does not allow us to precise the type of errors being produced by the translator. Therefore, besides the automatic evaluation of translations, some semi-automatic error analysis has also been done. In the works described in Popović and Ney (2006) and Popović et al (2006), errors in an English-Spanish statistical MT system were analysed with respect to their morphological and syntactic origin, and revealed problems in specific areas of inflectional morphology and syntactic reordering (Kirchhoff et al, 2007). A graphical user interface

[11] http://corpus.leeds.ac.uk/mellange/about_mellange.html

that automatically calculates various error measures for translation candidates and thus facilitates manual error analysis is presented in Niessen et al (2000).

In what concerns manual error identification, Bojar (2011) carried out a manual evaluation of four systems: Google, PC Translator[12], TectoMT[13] and CU-Bojar (Bojar et al, 2009). In his work, Bojar used two techniques of manual evaluation to identify error types of the previously mentioned MT systems. The first technique is called "blind post-editing" and consists of an evaluation performed by two people, separately. The first annotator receives the system output and has to correct it producing an edited version; meanwhile the second annotator gets the edited version, the source and the reference translation, and judges if the translation is still acceptable. The second technique used was the manual annotation of the errors using a taxonomy inspired by Vilar et al (2006).

A similar work is presented in Condon et al (2010), but with translations to and from English to Iraqi Arabic. Errors were annotated both as "Deletions", "Insertions" and "Substitutions" for morphological classes and types of errors following a similar taxonomy as the one proposed by Vilar et al (2006).

Also, in Fishel et al (2012) a collection of translation errors annotated corpora is presented, consisting of automatically produced translations and their detailed manual analysis[14]. Using the collected corpora, the authors evaluated two available state-of-the-art methods of MT diagnostics and assess: Addicter (Zeman et al, 2011)[15] and Hjerson (Popović and Hermann, 2011)[16]. Addicter is an open-source tool that uses a method explicitly based on aligning the hypothesis and reference translations to devise the various error types.

The Framework for Machine Translation Evaluation (FEMTI)[17] is a tool created to help people that evaluate MT systems. FEMTI has two classifications incorporated: the first one consists of characteristics of the contexts where the MT systems can be applied. The second one lists the MT system's characteristics, as well as the metrics proposed to measure them. People that use this framework have to specify the intended context for the MT system in the first classification and submit. In return, the FEMTI proposes a set of characteristics that are important in that particular context, using its embedded knowledge base. All the characteristics and evaluation metrics can be changed. After this task is completed, the evaluators can print the evaluation plan and do the evaluation.

To conclude, we should mention the work described in Secară (2005), which presents a survey on state-of-the-art translation evaluation methods, but on a much more linguistically oriented approach, where the focus of most of the analysed frameworks is on annotation schemes and error weighing for assessing the quality of a translated text, and on including post-editing feedback from human experts in error reductions and translation improvements.

[12] `http://langsoft.cz/translatorA.html`
[13] `http://ufal.mff.cuni.cz/tectomt`
[14] `http://terra.cl.uzh.ch/terra-corpus-collection.html`
[15] `https://wiki.ufal.ms.mff.cuni.cz/user:zeman:addicter`
[16] `http://www.dfki.de/~mapo02/hjerson`
[17] `http://www.issco.unige.ch:8080/cocoon/femti/st-home.html`

3 Taxonomy

Error identification is not always a straight-forward task. Not all errors are easily localizable: some are diffused throughout the sentence or larger units of text that contains them (James, 1998). Underlying the identification problem, remains the problem of their classification. Our taxonomy classifies errors in terms of "the linguistic item which is affected by the error" (H. Dulay and Krashen, 1982). Thus, the coarsish categories – **Orthography**, **Lexis**, **Grammar**, **Semantic**, and **Discourse** – indicate the language level where the error is located. In the following sections we explain each one of these categories and specify the subcategories of the linguistic units where the error occurs. This description is illustrated with errors resulting from EN to EP translations. As usual, each error is identified with an asterisk, which is placed before the error expression.

3.1 Orthography level

`Orthography` level errors include all the errors concerning misuse of punctuation and misspelling of words. We divide `orthography` level errors into three types: `punctuation`, `capitalization` and `spelling`. Each incorrect use of punctuation represents a `punctuation` error.

Example: `Punctuation error`
EN: *green tea*
EP: *chá*, verde*
Correct translation: *chá verde*

A `capitalization` error occurs when there is an inappropriate use of capital letters (for instance, the use of a small caption in the first letter of a proper noun). In the example above, the English sentence is correct, as the pronoun *I* is always spelt with a capital letter. Meanwhile, the Portuguese sentence does not have a subject (it is not expressed, but was previously mentioned) and probably because of this the verb was spelt with a capital letter.

Example: `Capitalization error`
EN: *... on time, I can console myself...*
EP: *... a tempo, *Posso consolar-me...*
Correct translation: *... a tempo, posso consolar-me...*

Finally, a `spelling` mistake concerns the substitution, addition or deletion of one or more letters (or graphic accent) to the orthography of a word.

Example: `Spelling error`
EN: *Basilica of the Martyrs*
EP: *Basílica dos *Mátires*
Correct translation: *Basílica dos Mártires*

Although a `capitalization` error could be considered a `spelling` mistake, we opted to provide both categories, and define them at the same abstraction level. This is due to the fact that if a capitalization error is common in natural language processing tasks, such as Automatic Speech Recognition and MT, a spelling mistake is not, as usually systems are trained with texts that do not have many spelling errors (news, parliament sessions, etc.). On the other hand, if we consider a human translation, spelling mistakes tend to be frequent, but capitalization errors are rare[18]. For this reason, we have decided to keep both type of errors in the taxonomy, so that both human an machine translations errors could be covered in it.

3.2 Lexis level

Under this category we have considered all errors affecting lexical items. It should be clear that, contrary to spelling errors that respect the characters used within a word, lexis errors concern the way each word, as a whole, is translated. Thus, the following types of errors at the `lexis` level are taken into account: `omission`, `addition` and `untranslated`. Moreover, `omission` and `addition` errors are then analysed considering the type of words they affect: a) **content words** (or lexical words), that is, words that carry the content or the meaning of a sentence (such as as nouns (*John*, *room*) or adjectives (*happy*, *new*)); b) **function words** (or grammatical words), that is, words that have little lexical meaning, but instead serve to express grammatical relationships with other words within a sentence (examples are, for instance, prepositions (*of*, *at*) and pronouns (*he*, *it*, *anybody*)).

`Omission` errors happen when the translation of a word present in the source text is missing in the resulting translation; an `addition error` represents the opposite phenomenon: the translation of a word that was not present in the source text and was added to the target text.

Example: `Omission` **error (content word)**
EN: *In his inaugural address, Barack Obama*
EP: *No seu * inaugural, Barack Obama*
Correct translation: *No seu discurso inaugural, Barack Obama*

Example: `Omission` **error (function word)**
EN: *In India*
EP: *Em Índia*
Correct translation: *Na Índia* (*Na* is the contraction of the proposition *em* and the article *a*, which was missing in the translation)

In the first example, the word *address*, a content word, was not translated and so it was missing from the sentence in Portuguese. In the second example, the missing word is a pronoun (function word). The country *India*, in Portuguese is always preceded by a definite article that in this case is missing from the translation output.

Example: `Addition` **error (content word)**
EN: *This time I'm not going to miss*
EP: *Desta vez *correr não vou perder*

[18] Although in some languages they can be more frequent, as for instance in German, where all nouns are spelled with capital letter, which can be a problem for foreign students that do not have this particularity in their mother tongue.

Correct translation: *Desta vez não vou perder*

Example: `Addition` **error (function word)**
EN: *highlights the work*
EP: **Já destaca-se o trabalho*
Correct translation: *destaca-se o trabalho*

These last two examples concern the addition of words to the translation output. In the first sentence, the translation engine added the verb *correr* (run) to the Portuguese sentence. Literally translating, the sentence was translated to *This time run I'm not going to miss.*. In the second example, the added word was a function word, the adverb *já* (alredy). In this case, the sentence was roughly translated to *Already highlights the work*. This not a grammatically or semantically wrong sentence, the only problem is that the word *already* was not on the source text.

Besides omitting or adding words in the translation, one other situations can occur: a word is not translated (`untranslated`).

An `untranslated` error situation is very common in MT, because when the engine cannot find any translation candidate to a given source word, an option is to copy it to the translation output. This option is frequently used as it results successfully if, for instance, a proper noun is not to be translated.

Example: `Untranslated` **errors**
EN: *in the world of botany*
EP: *no mundo da *botany*
Correct translation: *no mundo da botânica*

In this example, the translation engine did not have a translation for the word *botany*, so this word was simply copied to the output sentence.

3.3 Grammar level

`Grammar` level errors are deviations in the morphological and syntactical aspects of language. On this level of analysis we identified two types of errors: `misselection` errors and `misordering` errors.

`Misselection` errors are morphological misformations that the words may present, occurring on the grammatical level. This is the case of problems at word class level (for instance, an adjective is needed, but the translation engine returns a noun, instead), and at verbal level (tense and person). Errors of agreement (gender, number, person), and in contractions (between prepositions and articles) also fall into this type of error. When we have more than one of these problems in the same word we called it blend.

Example: `Misselection` **error (word class)**
EN: *world*
EP: *mundial* (*worldwide*)
Correct translation: *mundo*

Example: `Misselection` **error (verb level: tense)**
EN: *Even though this is a long list*

EP: *Mesmo que esta *é uma longa lista* (*é* (*to be*) should be in the subjunctive and not in the indicative mood)
Correct translation: *Mesmo que esta seja uma longa lista*

Example: `Misselection` **error (verb level: person)**
EN: *Theater-goers can discover*
EP: *As pessoas que vão ao teatro *pode descobrir* (the correct form of the verb is in the the third person plural and not in the third person singular)
Correct translation: *As pessoas que vão ao teatro podem descobrir*

Example: `Misselection` **error (verb level: blend)**
EN: *If I go to see the Dario Fo play*
EP: *Se *vai ver a peça de Dário Fo.* (*ir* (*to go*) should be in the conditional and not in the indicative mood and on the first person singular not the third person.)
Correct translation: *Se for ver a peça de Dário Fo.*

Example: `Misselection` **error (agreement: gender)**
EN: *The German artist Thomas Schutte*
EP: **(A artista alemã) Thomas Schutte* (in Portuguese, like all morphologically rich languages, the pronoun and adjective have to agree in gender and number with the noun, in this case masculine and singular, not feminine)
Correct translation: *O artista alemão Thomas Schutte*

Example: `Misselection` **error (agreement: number)**
EN: *moral skills*
EP: *capacidades *moral* (both the adjective and noun have to be in the plural form)
Correct translation: *capacidades morais*

Example: `Misselection` **error (agreement: person)**
EN: *learn from our failures*
EP: *aprender com os *vossos fracassos* (the use of the possessive pronoun *vossos* is grammatically correct, but it is not in the correct person)
Correct translation: *aprender com os nossos fracassos*

Example: `Misselection` **error (agreement: blend)**
EN: *funky clothes shops*
EP: *lojas *simpático de roupa* (the adjective is not in the correct gender and number)
Correct translation: *lojas simpáticas de roupa*

`Contraction` problems are typical of Romance languages, such as Portuguese, but also of languages such as German, as many prepositions, for instance *em*, are compulsory contracted with an article, for example *na* that results from the contraction of the preposition *em* (*in*) and the article *a* (*a*)[19].

[19] We should not confuse `Omission` errors of a function word with `Misselection` errors (contraction). In the first case, in the phrase *na (em + a) Indía* (*in India*), the article *a* is missing, so we have an `Omission` error. Meanwhile, if we had a `contraction` problem, the sentence would be *em a India*, both preposition and article were correctly selected but they were not contracted as they should be (*em* + *a* = *na*).

Example: `Misselection` **error (contraction)**
EN: *in an environment*
EP: *em um ambiente*
Correct translation: *num ambiente*

Finally, `Misordering` errors are related with syntactical problems that the sentences may demonstrate. We should point out that a good translation is not only selecting the right forms to use in the right context, but also to arrange them in the right order. In Portuguese, certain word classes such as adverbials and adjectives seem to be especially sensitive to misordering.

Example: `Misordering` **error**
EN: *A person is wise.*
EP: *Uma pessoa sábia *é.* (*A person wise *is.*)
Correct translation: *Uma pessoa é sábia.*

3.4 Semantic level

By semantic errors we understand problems that regard the meaning of the words and subsequent wrong word selection. We have individuated three different types of errors: `confusion of senses`, `wrong choice`, `collocational error` and `idioms`.

`Confusion of senses` is the case of a word that was translated into something representing one of its possible meanings, but, in the given context, the chosen translation is not correct.

Example: `Confusion of senses` **errors**
EN: *the authentic tea set that includes a tray, teapot and glasses* (*glasses* means spectacles, but it is also the plural of the noun *glass*)
EP: *um autêntico jogo de chá que inclui bandeja, bule e *óculos* (*the authentic tea set that includes a tray, teapot and spectacles*)
Correct translation: *um autêntico jogo de chá que inclui bandeja, bule e copos*

In what concerns `wrong choice` errors, they occur when the wrong word, without any apparent relation, is used to translate a given source word.

Example: `Wrong choice` **errors**
EN: *in the same quarter*
EP: *no mesmo *histórica* (*in the same *historical*)
Correct translation: *no mesmo bairro*

We should not confuse `Wrong Choice` with `Confusion of senses`, an example of the first case is, for instance, the translation of *care* as *conta* (check), there is no semantic relation between these two words. As for the translation of *glasses* as *óculos* (glasses) is a predictable `Confusion of senses`, as the English word *glasses* can be translated into two different words in Portuguese: glasses to drink (*copos*) and glasses to see (*óculos*).

Collocations are the other words any particular word normally keeps company with (James, 1998). They have a compositional meaning, contrary to idioms, but the selection of their constituents is not semantically motivated. `Collocational` errors could be considered an instantiation of the previous error, but we have decided to take them into consideration separately. This decision was made because in the case of `Confusion of senses` errors, we account for single words misusage; meanwhile `Collocational` errors occur on blocks of words.

Example: `Collocational errors`
EN: *high wind*
EP: *vento alto* (literally means *tall wind*)
Correct translation: *vento forte*

`Idiomatic` errors concern errors in idiomatic expressions that the system does not know and translates as regular text. These expressions cannot be literally translated as their meaning it is not literal and, in many cases, the equivalent expression in the target language is very different.

Example: `Idioms`
EN: *kick the bucket*
EP: *dar um pontapé ao balde*
Correct translation: *esticou o pernil* (idiomatic expression that means *to die*)

3.5 Discourse level

By `Discourse` level errors, we understand discursive options that are not the most expected, but still are not errors. We consider three different situations at the `Discourse` level: `style`, `variety` and `should not be translated`. In all this cases, the meaning is preserved (thus, they are not semantic errors), but the chosen word is not the best choice.

`Style errors` concern a bad stylistic choice of words when translating a sentence. A typical example is the repetition of a word in a near context, where a synonym should have been selected.

Example: `Style errors`
EN: *permission to be allowed to improvise*
EP: *autorização para ser autorizado a improvisar* (*permission to be permitted to*)
Correct translation: *permissão para ser autorizado a improvisar*

`Variety` errors cover the cases when the target of the translation is a certain language, but instead lexical or grammatical structures from a variety of that language are used. This is what happens, for instance, when the target of a translation is EP and Brazilian Portuguese (BP) is returned, which is very common in Google translations. With `Variety errors`, this taxonomy is then able to capture this phenomenon.

Example: `Variety errors`
EN: *in his speech*
EP: *em seu discurso*
Correct translation: *no seu discurso* (in EP, we need an article before the possessive pronoun (*seu*) that, in this case, is contracted with the preposition *em* (*em* + *o* = *no*))

Under the `should not be translated` category we have considered all the word's sequences in the source language that should not be translated in the target language. In this particular case, we can find, for instance, books or film titles. In this example, we have the name of a Portuguese play, in the text in English, the tittle was left in Portuguese, but the engine tried to translate it and only added errors.

Example: `Should not be translated errors`
EN: *Havia um Menino que era Pessoa*
EP: *Havia hum Menino era Opaco Pessoa*
Correct translation: *Havia um Menino que era Pessoa*

3.6 General Overview

In Figure 1 we resume the taxonomy previously presented.

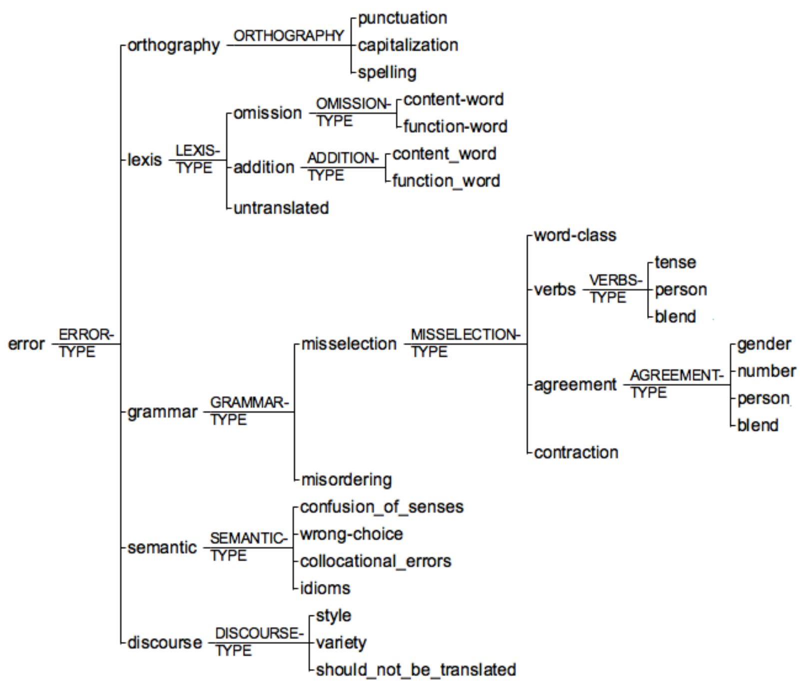

Fig. 1 Taxonomy

3.7 Comparison with other taxonomies

In this section we compare our taxonomy with the ones created for MT and described in Section 2, namely the ones presented in Bojar (2011) and Vilar et al

(2006), as well as with the one traditionally used for classifying students' errors, also described in Section 2 and detailed in H. Dulay and Krashen (1982). From now on, we will call these taxonomies **Bojar**, **Vilar** and **Dulay**, respectively. Table 1 summarizes the following discussion.

Error's types		**Bojar**	**Vilar**	**Dulay**
Orthography	Punctuation	✓	✓	✓
	Capitalization	✘	✘	✘
	Spelling	✘	✘	✓
Lexis	Omission	✓	✓	✓
	Addition	✓	✓	✓
	Untranslated	✓	✓	✓
Grammar	Word Class	✓	✘	✓
	Verbs	✓	✓	✓
	Agreement	✘	✓	✓
	Contraction	✘	✘	✓
	Misordering	✓	✓	✓
Semantic	Confusion of senses	✓	✓	✓
	Wrong choice	✓	✓	✓
	Collocational errors	✘	✘	✘
	Idioms	✘	✓	✓
Discourse	Style	✘	✓	✓
	Variety	✘	✘	✘
	Should not be translated	✘	✘	✘

Table 1 Comparision with other taxonomies

In what regards `orthography` level errors, **Bojar** and **Vilar** taxonomies can only consider `punctuation` errors. On the contrary, **Dulay**'s taxonomy, besides `punctuation` errors also considers `spelling` errors. As previously said, our taxonomy also considers `capitalization` errors.

Concerning `lexis` errors all the taxonomies agree on the proposed different types of errors, but there are significant differences in what respects `grammar` errors: **Vilar** does not contemplate `word class` errors, mentioned by all the other taxonomies, and, at the level, it only classifies `gender` and `number`, while we found important to add other types of agreement, as they are significant for Portuguese and Romance languages in general; **Bojar** does not talk about `agreement` problems, but **Dulay** does take this type of error into consideration. `Contraction` errors are mentioned by none of these three taxonomies. We hypothesize that this happens because most of the taxonomies are built for English and this phenomena is not compulsory in English, as contractions are an informal option of language usages. However, contractions are in several cases obligatory for Portuguese and languages like Italian, German, Spanish or French, for this reason we consider that `contraction` errors must be contemplated in an error's taxonomy.

Regarding `semantic` errors, **Vilar** does not mention `collocational` errors and **Bojar** does not mention `idiomatic` nor `collocational` errors. **Dulay** mentions all these types of `semantic` errors. All of the previously mentioned authors report the existence of `wrong choice` of word errors. At the `discourse` level, `style` mistakes are mentioned by Vilar et al (2006), but not by Bojar (2011). **Dulay** only takes into consideration style mistakes, as **Dulay**'s taxonomy was conceived to assess human errors, not all the categories used in machine translation could have a direct equivalent in an students typology of mistakes. `Variety` errors are assessed only by our taxonomy. As previously mentioned, this type of error is very frequent in Google's translations of EP. However, the same thing happens between American English and British English, although this type of error is not considered in taxonomies built for English.

3.8 Language dependent and independent errors

After the discussion of our purposed taxonomy it is important to differentiate between the errors that could happen in any machine translation task and that are independent of the source and output language, and the errors that are language dependent and that may not occur in every language. On Table 2 we present a resume of the following discussion.

Error's types		**Language independent**
Orthography	Punctuation	✓
	Capitalization	✘
	Spelling	✓
Lexis	Omission	✓
	Addition	✓
	Untranslated	✓
Grammar	Word Class	✘
	Verbs	✘
	Agreement	✘
	Contraction	✘
	Misordering	✘
Semantic	Confusion of senses	✓
	Wrong choice	✓
	Collocational errors	✓
	Idioms	✓
Discourse	Style	✓
	Variety	✘
	Should not be translated	✓

Table 2 Language independent errors.

We can not assure that every existing language has `punctuation`, but we know that a great majority of them have. We should point out that the punctuation symbols may be different. In Greek, the question mark is written as the English semicolon and in Spanish an inverted question mark is used at the beginning of a question and the normal question mark is used at the end.

`Capitalization` errors should be considered language specific, as languages like Chinese, Arabic or Korean do not have capital letters.

Considering `spelling` mistakes, every written down languages has a standard orthography, and any misuse of these rules may be marked as a spelling mistake.

To what concerns `lexis` errors, `omissions`, `additions` and `untranslated` words can be present in any automatic translation, independently of the language. Regarding grammatical problems, this category of mistakes is language specific and not all of the `grammatical` error types make sense in every language. According to Keenan and Stabler (2010) 'different languages do have non-trivially different grammars: their grammatical categories are defined internal to the language and may fail to be comparable to ones used for other languages. Their rules, ways of building complex expressions from simpler ones, may also fail to be isomorphic across languages.'

As De Saussure (1916) defended, language is ambiguous and polissemic by definition. By this we mean that in every language there are `semantic` problems that can arise in an automatic translation. For instance, any idiomatic expression has a non-literal meaning that can not always be captured by a literal translation. Information about the context of use is necessary for it to be well interpreted and translated.

Concerning `discourse` errors, `style` errors may occur in every language, as different social contexts require an appropriate discourse. To what concerns `variety`, not all languages have a variety, like American English and British English or European Portuguese and Brazilian Portuguese, so this category should only be used for languages that have a different variety.

Finally, words that `should not be translated` and that should be kept in the language of the source language is a problem that is language independent.

4 Experimental Setup

In this section we briefly describe the corpora and the tools we have used in our experiments; we also present the annotation agreement resulting from the annotation of the translation errors (according with the proposed taxonomy) in each corpus.

4.1 Corpora

As previously said, the error analysis was carried out on a corpus of 750 sentence pairs, composed of:

- 250 pairs of sentences taken from TED talks – from now on the TED corpus;
- 250 pairs of sentences taken from the UP magazine from TAP – from now on the TAP corpus;
- 250 pairs of questions taken from a corpus made available by Li and Roth (2002), from the TREC collection – from now on the Questions corpus.

The TED corpus is composed of TED talk subtitles and corresponding EP translations. These were created by volunteers and are available at the TED website. As we are dealing with subtitles (and not transcriptions), content is aligned to fit the screen, and, thus, some pre-processing was needed. Therefore, we manually connected the segments in order to obtain parallel sentences.

The TAP corpus is constituted of 51 editions of the bilingual Portuguese national airline company magazine, divided into 2 100 files for EN and EP. It has almost 32 000 aligned sentences and a total of 724 000 Portuguese words and 730 000 English words.

The parallel corpus of Questions (EP and EN) consists of two sets of nearly 5 500 plus 500 questions each, to be used as training/testing corpus, respectively.

Details on its translation and some experiments regarding statistical machine translation of questions can be found in Costa et al (2012).

Some examples of sentences from these corpora can be found in Table 3.

TED	The publisher bears no responsibility for return of unsolicited material and reserves the right to accept or reject any editorial and advertising material. No parts of the magazine may be reproduced without the written permission of up. The opinions expressed in this magazine are those of the authors and not necessarily those of the auditor.
TAP	They're the things you would expect: mop the floors, sweep them, empty the trash, restock the cabinets. It may be a little surprising how many things there are, but it's not surprising what they are.
Questions	Who developed the vaccination against polio? What is epilepsy? What year did the Titanic sink? Who was the first American to walk in space?

Table 3 Examples of sentences from the corpora.

The Questions corpus has short sentences and most of them start by an interrogative pronoun. The TAP corpus presents more complex grammatical structures when compared with the TED corpus, which is influenced by its semi-spontaneous nature (some previous preparation is involved). This difference may be observed because written language tends to be more complex and intricate than speech, with longer sentences and many subordinate clauses. Spoken language tends to be full of repetitions, incomplete sentences, corrections and interruptions, which sometimes result in agrammatical sentences.

On Table 4, there are some details on the number of sentences, tokens and average number of tokens by sentence that were translated. By token we understand a string of characters delimited by a white space. Therefore, not only words, but also punctuation, are tokens.

Dataset	Language	Sentences	Tokens	Average sentence length
TAP	EN	250	4 868	19.47
	EP	250	5 521	22.08
TED	EN	250	3 346	13.38
	EP	250	3 894	15.58
Questions	EN	250	1 856	7.42
	EP	250	2 048	8.19

Table 4 Data used in the error analysis.

Finally, on Table 5 we can see the number of tokens per translated dataset, and the average number of tokens per sentence for each dataset.

Dataset	TAP	TED	Questions
Online-S	5 725 22.90	3 855 15.42	1 955 7.82
Online-G	5 623 22.49	3 956 15.82	2 030 8.12
Moses-PSMT	5 522 22.09	3 730 14.92	2 068 8.27
Moses-HSMT	5 507 22.03	3 759 15.02	2 059 8.24

Table 5 Number of tokens of the translated corpora on the first line, and, on the second line, the average number of tokens per sentence for each dataset.

4.2 Systems and tools

4.2.1 Machine Translation Systems

We tested four different systems in our evaluation: two mainstream online translation systems (Google Translate (statistical) and Systran (hybrid)), and two in-house MT systems. The online systems were run as they were[20] and we will denote them as Online-G and Online-S, respectively. The in-house systems were trained using Moses, and the two popular models: the phrase-based model (Koehn et al, 2007) and the hierarchical phrase-based model (Chiang, 2007), which we will denote as Moses-PSMT and Moses-HSMT, respectively. Both systems share the same training corpora, comprised of approximately 2 million sentence pairs from Europarl (Koehn, 2005). As for the in-domain corpora, we gathered the remaining sentence pairs for the TAP, Questions and TED domain after removing the held-out data, and added these into the training corpora. These contained 4 409, 8 904 and 78 135 sentence pairs, respectively. In total, there were 56 million tokens in English (27.32 tokens per sentence) and 58 million tokens (28.28 tokens per sentence) for Portuguese in the training set[21].

The model was built by first running IBM model 4, with Giza++ (Och and Ney, 2003), and bidirectional alignments were combined with the grow-diag-final heuristic, followed by the phrase extraction (Ling et al, 2010) for the Moses-PSMT model and rule extraction (Chiang, 2007) for the Moses-HSMT model. The parameters of the model were tuned using MERT (Och, 2003) and we used 1 000 sentence pairs from each of the domains for this purpose. The statistics of the data used are detailed in Table 6. The splits were chosen chronologically in the order the sentences occurred in the dataset. We can see that the majority of the training data is out-of-domain (Europarl), and a relatively small in-domain parallel dataset was used. Translations were also detokenized using the Moses detokenizer, and capitalized using the Portuguese capitalizer described in Batista et al (2007).

[20] Translated on 22/10/2014
[21] Tokenized using the default moses tokenizer

Dataset	Train Sentences	Tuning Sentences	Test Sentences
TAP	4 409	1 000	250
TED	78 135	1 000	250
Question	8 914	1 000	250
Europarl	1 960 407	0	0

Table 6 Data used for training, tuning and testing the MT models.

4.2.2 UAM CorpusTool

Our corpus was annotated using UAM CorpusTool[22], a state-of-the-art environment for annotation of text corpora (see Figure 2).

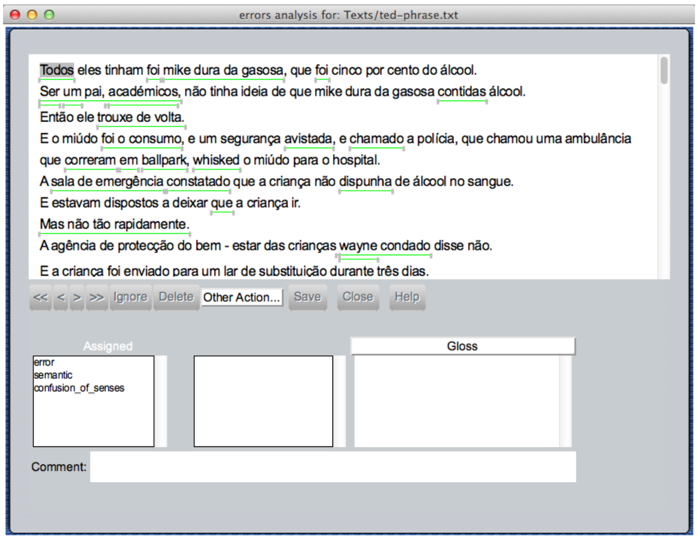

Fig. 2 UAM Corpus Tool

This tool allows the annotation of multiple texts using the annotation schemes previously designed. The annotation is simply done by swiping some text (clicking down and dragging to the end of the segment) and then indicating the features that are appropriate for that segment. This tool also supports a range of statistical analyses of the corpora, allowing comparisons across subsets.

4.3 Annotation agreement

For agreement purposes we compared the answers from two annotators: the linguist that developed the taxonomy and annotated the presented corpora, and another annotator with no formal linguistic instruction. The latter was given an explanation about the different types of errors on the taxonomy, and a set of annotation

[22] http://www.wagsoft.com/CorpusTool

guidelines. Both annotators identified and classified the errors on a total of 300 sentences: 25 sentences per dataset translated by the four MT engines. Table 7 shows the agreement between the two annotators using Cohen's kappa.

	Questions	TAP	TED
Localization	0.9717	0.9622	0.9670
L1	0.8295	0.9441	0.9140
L2	0.9216	0.9763	0.9434
L3	0.8223	0.9622	0.9662
L4	1	0.9554	0.9768

Table 7 Inter-annotator agreement

The agreement is first computed in terms of error localization, that is, whether the annotators agree where the error is placed in the sentence. Errors can be found not only in words or punctuation, but also between words (since words or punctuation marks might be missing). For each word, punctuation mark and space, we measure the agreement on a binary decision regarding the existence or non-existence of an error.

Then, we take all cases where annotators agree that there is an error, and check if they also agree on the classification of the error, considering the first level (L1) of the taxonomy (`Orthography`, `Lexis`, `Semantic`, `Grammar` and `Discourse`). We repeat the same process for the second, third and fourth levels, where we gather all cases where the annotators agree on the previous level (it is possible that the annotators agree on a certain level and agree or not on the next one), and compute the agreement coefficient. As Table 7 shows, the agreement on the identification of the errors in the sentences is high for all three data sets, specially for the corpus of questions.

5 Error Analysis

In the following we analyze the errors carried out by Online-S, Online-G, Moses-PSMT and Moses-HSMT, according to the proposed taxonomy and in the different scenarios.

5.1 Preliminary remarks

Before we start analysing our results, there are two important issues that need to be discussed:

- We will present our results as the number of errors per dataset, but it should be clear that some words have more than one error, namely two. We have calculated the cases of words with two errors in the total of number of errors, and only between 3.05% and 11.18% of the words have two errors. Except for Moses-HSMT that in the Question corpus has 16.16% of words with two errors;

- A straightforward comparisons of the errors' types between systems is only possible at the `lexis` level. This is due to the fact that although in some situations a word may have two different error tags (for instance `misordering` and `capitalization`), words that remain untranslated or that were omitted in the target, will never have errors at `grammar` or `semantic` level. Thus, we cannot use the number of errors to compare systems after the `lexis` level. For instance, consider that the English sentence *He was sick yesterday.* was translated as **Ele doente ontem.* (verb omitted) by system A, and as **Ele está doente ontem* by system B (verb in the wrong form). Then, system A will have a `lexis` type error and system B a `grammar` level error. However, we cannot say that system B has more grammatical errors that system A. That is, the number of errors can be used by each system mainly as an indicator of what the system is doing wrong.

5.2 General overview

Table 8 summarizes the percentage of errors by translation scenario relative to the number of tokens per corpus.

System	TAP	TED	Questions
Online-S	21%	20%	25%
Online-G	10%	14%	13%
Moses-PSMT	16%	19%	20%
Moses-HSMT	18%	21%	19%

Table 8 Percentage of errors.

From these results, we can observe that:

- For the Online-S system the corpus of Questions was the most problematic document (25% of errors). Although the syntactical form of questions is usually very simple (for instance, *What is epilepsy?*), there are problems choosing the right interrogative pronoun. For instance, Online-S translates the sentence *What is the population of Nigeria?* into **Que é a população da Nigéria?*, instead of *Qual é a população da Nigéria?*. That is, *Que* should have been translated as *Qual*;
- Although the TAP magazine is constituted of long sentences[23], is was the corpus that caused less problems for the majority of the systems (10%, 16% and 18%, for Online-G, Moses-PSMT and Moses-HSMT, respectively).

Table 9 summarises the number of errors found for each error type.

In the next sections, we will discuss each specific type of error.

[23] For instance, the following sentence is in the TAP corpus: *If it's Saturday, there's a play at Teatro da Trindade called Havia um Menino que era Pessoa, where theatre-goers can discover the verses the poet wrote for his nephews and nieces.*

Orthography	TAP	TED	Questions	Errors (total)
Online-S	32	13	24	69
Online-G	19	12	3	34
Moses-PSMT	145	22	51	218
Moses-HSMT	148	23	52	233
Lexis	**TAP**	**TED**	**Questions**	**Errors (total)**
Online-S	441	223	219	883
Online-G	164	133	83	380
Moses-PSMT	247	223	136	606
Moses-HSMT	315	257	128	700
Grammar	**TAP**	**TED**	**Questions**	**Errors (total)**
Online-S	312	227	90	629
Online-G	157	175	72	404
Moses-PSMT	251	269	129	649
Moses-HSMT	288	296	129	713
Semantic	**TAP**	**TED**	**Questions**	**Errors (total)**
Online-S	357	288	138	783
Online-G	118	135	81	334
Moses-PSMT	211	182	93	486
Moses-HSMT	225	184	80	489
Discourse	**TAP**	**TED**	**Questions**	**Errors (total)**
Online-S	78	39	17	134
Online-G	79	84	23	186
Moses-PSMT	21	7	9	37
Moses-HSMT	19	10	7	36

Table 9 System's error types.

5.3 Lexis Level Errors

According to Table 10, Moses-PSMT and Moses-HSMT performed considerably worse than Online-G and Online-S on the number of untranslated words or expressions. Although `untranslated` words represent the minority of `lexis` errors, this clearly shows a direction on Moses developers research: the translation of unknown words (that is, words never seen during training).grammar

Regarding `addition` and `omission` errors, the system with more errors of this kind was Online-S (388 and 449, respectively) and the system with less `addition` and `omission` problems was Online-G with 133 and 220, respectively. We should also mention that the majority of words that were added or elided were function words and not content words.

5.4 Grammar Level Errors

On the `grammar` level of errors, shown on Table 11, we have identified `misselection` and `misordering` errors (remember that `misselection` errors can affect verbs, agreements and contractions, while `misordering` include word order problems).

In what concerns `misordering` errors, the smaller number of errors could be explained by the common feature shared by English and Portuguese: the subject-

Omission	TAP	TED	Questions	Errors (total)
Online-S	184	118	147	449
Online-G	104	71	45	220
Moses-PSMT	103	104	63	270
Moses-HSMT	147	122	47	316
Addition	**TAP**	**TED**	**Questions**	**Errors (total)**
Online-S	236	92	60	388
Online-G	54	53	26	133
Moses-PSMT	100	84	55	239
Moses-HSMT	121	104	60	285
Untranslated	**TAP**	**TED**	**Questions**	**Errors (total)**
Online-S	21	13	12	46
Online-G	6	9	12	27
Moses-PSMT	44	35	18	97
Moses-HSMT	47	31	21	99

Table 10 Lexis Errors

Misordering	TAP	TED	Questions	Errors (total)
Online-S	112	55	51	218
Online-G	36	33	31	100
Moses-PSMT	66	54	53	173
Moses-HSMT	96	97	56	249
Misselection	**TAP**	**TED**	**Questions**	**Errors (total)**
Online-S	200	172	39	411
Online-G	121	142	41	304
Moses-PSMT	185	215	76	476
Moses-HSMT	196	199	73	464

Table 11 Grammar Errors

verb-object (SVO) structure[24]. However, although in a syntactic point of view English and Portuguese have in common some aspects, there are also differences. For instance, considering the order of the noun phrase (usually a noun, pronoun, or other noun-like word (nominal), which is optionally accompanied by a modifier such as adjectives), in English the correct sequence is Adjective + Noun; meanwhile, in Portuguese, the usual order is the opposite, although in certain contexts the order Adjective + Noun is also possible. This idiosyncratic aspect of language may explain errors, like **fagrammarvorito artista* (*favourite artist*), **permanente colecção* (*permanent collection*), **alemã artista* (*german artist*), **artística carreira* (*artistic career*). Another structure that might influence word order is the position of the direct-object and indirect-object pronouns. In Portuguese, the rule says that these pronouns should be placed after the verb, but there are many exceptions, being an example of this the sentence wrongly translated by Online-G *para comprá-lo* (to buy it). In this case of infinitive construction with the preposition *para* (to), the pronoun should be placed before the verb *para o comprar* (to buy it). Meanwhile, in English, it should be verb + pronoun (*to buy it*).

[24] An SVO is a sentence structure where the subject comes first, the verb second, and the object third, and languages may be classified according to the dominant sequence of these elements. SVO is one of the most common order in world languages.

Considering `misselections` errors, Table 12 summarises the associated errors.

Word class	TAP	TED	Questions	Errors (total)
Online-S	40	20	12	72
Online-G	6	4	5	15
Moses-PSMT	20	31	9	52
Moses-HSMT	23	29	4	56
Verbs	**TAP**	**TED**	**Questions**	**Errors (total)**
Online-S	55	69	16	140
Online-G	38	79	20	137
Moses-PSMT	49	94	25	168
Moses-HSMT	54	75	24	153
Agreement	**TAP**	**TED**	**Questions**	**Errors (total)**
Online-S	96	71	7	174
Online-G	67	50	10	127
Moses-PSMT	107	89	42	238
Moses-HSMT	106	91	43	240
Contraction	**TAP**	**TED**	**Questions**	**Errors (total)**
Online-S	9	12	4	25
Online-G	10	9	6	25
Moses-PSMT	8	1	0	9
Moses-HSMT	9	4	1	14

Table 12 Misselections

The most common errors were `agreement` errors. We should point out that in Portuguese, according to Cunha and Cintra (1987), the general rule for adjectives is that they agree in gender and number with the noun they modify. In English, the agreement between the adjective and the closest noun is restricted to the words *this* and *that* (as well as *these* and *those*), as these are the only that have separate forms for singular and plural. This structural difference between languages explains many translation errors, like *a wise man* being translated into Portuguese as **um sábio pessoa* (instead of *uma sábia pessoa*) by Moses-PSMT, and *the exhibition* translated as **o exposição* (instead of *a exposição*) by Online-G. In both cases the adjectives and articles have to agree in number and gender with the noun, but that does not happen, as we have a feminine noun (*pessoa*) with a masculine article, and a feminine noun (*exposição*) with a masculine article (*o*). Also looking at Table 12, we can notice that there were some problems producing the correct form of the required verb, specially on the TED talks. Portuguese has a variety of tenses, aspects, and moods, as well as constructions with auxiliary verbs that makes it more grammatically complex than, for example, English. For instance, in Portuguese the verb *estar* (*to be*) is used with the present gerund to indicate the present continuous aspect and the verb *ter* (*to have*) is used with the past participle for the perfect. English has a less complex tense system, and it is not a simple task, not even for a human translator, to find the correct correspondence between both languages tenses.

At last, we should take into consideration the type of `misselection` that had the less number of occurrences: `contractions`. This result was quite unexpected as this is another aspect of language where there is no congruence between English and Portuguese. In Portuguese contractions are in several cases compulsory. For instance, the preposition *de* (*of*) can be contracted with an article and become, for instance: *do* (*of* + masculine singular article), *da* (*of* + feminine singular article), *duns* (*of* + masculine plural article), *dumas* (*of* + feminine plural article). This language rule explains errors such as *em um* (*in* + masculine singular article) by Online-G and *por a* (*by* + feminine singular article) by Moses-PSMT (the correct forms are *uns* and *pela*, respectively).

5.5 Semantic and Discourse Errors

Now taking a closer look at Table 13, we observe that the `confusion of senses` error represents the majority of the `semantic` errors made by all engines.

Confusion of senses	**TAP**	**TED**	**Questions**	**Errors (total)**
Online-S	293	261	124	678
Online-G	94	118	57	269
Moses-PSMT	156	144	73	373
Moses-HSMT	179	154	67	400
Wrong choice	**TAP**	**TED**	**Questions**	**Errors (total)**
Online-S	47	19	13	79
Online-G	17	7	22	46
Moses-PSMT	45	27	18	90
Moses-HSMT	36	18	13	67
Collocational error	**TAP**	**TED**	**Questions**	**Errors (total)**
Online-S	11	5	1	17
Online-G	5	5	2	12
Moses-PSMT	5	8	1	14
Moses-HSMT	6	8	0	14
Idioms	**TAP**	**TED**	**Questions**	**Errors (total)**
Online-S	5	3	0	8
Online-G	2	5	0	7
Moses-PSMT	5	3	0	8
Moses-HSMT	4	4	0	8

Table 13 Semantic Errors

Many of the `confusion of senses` errors are due to prepositions as, for instance, *to* in Portuguese, which can have several translations (*para*, *a* or *de*, just to mention a few). The same happens with copular verbs, like *be* that in Portuguese can be translated into *ser*, *estar* or *ficar*. In the case of Moses-PSMT and Moses-HSMT, some of these errors can be linked with the Europarl nature. For instance, *ask* is translated as *auscultarem* (auscultate), *you* as *excelência* (excellency), *house* as *assembleia* (assembly) and *sitting* as *sessão* (sesson).

Considering `wrong choice` errors, these can produce translations with no apparent semantic explanation. An example of this is the translation of *understand* into *tradição* by Moses-HSMT.

Finally, `collocational` errors and `idioms` were not a significant problem.

In what concerns **discourse** errors, Table 14 summarises the observed number of errors.

Style	TAP	TED	Questions	Errors (total)
Online-S	9	8	0	17
Online-G	2	7	0	9
Moses-PSMT	4	2	2	8
Moses-HSMT	4	6	0	10
Variety	**TAP**	**TED**	**Questions**	**Errors (total)**
Online-S	31	23	6	60
Online-G	56	73	22	151
Moses-PSMT	0	0	0	0
Moses-HSMT	1	0	0	1
Should not be translated	**TAP**	**TED**	**Questions**	**Errors (total)**
Online-S	28	8	11	57
Online-G	20	3	1	24
Moses-PSMT	17	5	7	29
Moses-HSMT	14	4	7	25

Table 14 Discourse Errors

Clearly, in translating into EP (and not BP) is where Online-G performs worse, as there are many `variety` errors, resulting from the fact that this engine is translating EN into BP.

5.6 BLEU and METEOR scores

We performed an automatic MT evaluation using BLEU (Papineni et al, 2002) and METEOR[25] (Denkowski and Lavie, 2014), which uses other linguistic resources such as paraphrases. Results are presented on Table 15.

BLEU	TAP	TED	Questions
Online-S	17.94	19.66	32.98
Online-G	30.10	27.27	58.76
Moses-PSMT	45.23	26.92	42.97
Moses-HSMT	44.15	27.17	43.10
METEOR	**TAP**	**TED**	**Questions**
Online-S	38.59	39.51	51.76
Online-G	49.86	48.67	72.79
Moses-PSMT	58.84	45.98	62.02
Moses-HSMT	57.51	45.43	61.86

Table 15 BLEU and METEOR scores achieved by the translation systems when evaluated on each test dataset.

The MT system that had the lowest number of errors in the Questions corpus, was Online-G. This is also the system with the best BLEU (58.76) and METEOR scores (72.79). Both the human and the automatic evaluation metrics agree that the system with more words tagged as errors is Online-S.

[25] version 1.4

In the TED corpus, the best BLEU and METEOR scores are the ones from Online-G (27.27 and 48.67, respectively). According to the human evaluation this is also the system that has the lowest number of errors.

Interestingly, in the TAP corpus, we observe that our in-house SMT systems obtain a significantly better result in terms of BLEU and METEOR over Online-G, which is trained on more data, and outperforms our systems in all other datasets. However, we can observe that in terms of the total number of errors (as obtained by the human evaluation), system Online-G actually performs better than the in-house systems. In order to explain this inconsistency, we first tested if our trained systems were over-fitting the domain using the in-domain data. This was done by training systems using only the Europarl dataset. Results are shown in Table 16, where we observe that the BLEU and METEOR scores for both Moses systems, even though having dropped drastically, are still higher than for the system Online-G (see Table 15).

BLEU	**TAP**	**TED**	**Questions**
Moses-PSMT	37.50	26.59	38.60
Moses-HSMT	38.27	25.61	38.60
METEOR	**TAP**	**TED**	**Questions**
Moses-PSMT	52.68	44.76	57.81
Moses-HSMT	53.12	44.04	57.09

Table 16 BLEU and METEOR scores achieved by the translation systems trained online with the Europarl corpus.

This shows that the results are not caused by over-fitting the training data as the Europarl dataset is radically different from the TAP dataset. However, this could still be due to the tuning corpus, which is in-domain. We looked at the development corpora and noticed that there are many equivalent sentence pairs, such as menu items and general flight instructions, which are present in every issue of the magazine. Furthermore, many sentences are simply repetitive, such as, *Have a good flight.* and *Fancy a snack.* This happens with 53 sentence pairs (approximately 20%) in the TAP corpus, while there are only 5 and 2 (less than 1%) repeated sentences in the Questions and TED datasets, respectively. This allows the MT systems to tailor their output so that the translations of such content are as close as possible to the reference. This gives a large boost in the BLEU and METEOR scores as they are biased towards finding translations that are close to the reference and not by correctness. Still, it is an interesting result that even in such conditions, human evaluators find more errors in Moses-PSMT and Moses-HSMT than in Online-G. This shows the many shortcomings of these metrics, which rely on closeness to the reference or references rather than analysing their linguistic quality.

To confirm the hypothesis that the boost in scores is due to the common content on every issue of the magazine (such as flight instructions or menu items), we selected 50 sentences extracted from the magazine's main articles only. The recomputed METEOR and BLEU metrics on this subset are consistent with former experiments. As previously, Online-G (METEOR = 43.49; BLEU = 25.81) ranks first, while Moses-PSMT (METEOR = 35.31; BLEU = 22.14) and Moses-

HSMT (METEOR = 39.10; BLEU = 22.58) rank lower and Online-S ranks last (METEOR = 33.19; BLEU = 15.85). These scores are back in tune with human annotation, which assigns better translation results to Online-G than to both Moses systems.

6 Error gravity

Having just investigated how errors occur on four different systems and translation scenarios (Questions, TED and TAP datasets), we decided to analyze to what extent distinct error types impact translation quality. To accomplish that, we start with a subjective evaluation of the MT outputs, which consisted of ranking the four translations of each sentence (Section 6.1). Then, by relating this rank with our taxonomy, we are able to show how the presence of each error type reflects on quality (Section 6.2).

6.1 Ranking translations

Using the same set of 75 sentences that were used in the experiment reported in Section 4.3, we carried out an evaluation similar to the one proposed in Callison-Burch et al (2007). This task consists of presenting the annotator with the input sentence, the correct translation and all four MT outputs. He/she then decides on the order of translations based on his assessment of quality, ranking them from 1 (best) to 4 (worst). In our experiment, however, ties should only exist for translations that are exactly the same. This encouraged judges to make a decision even when facing tenuous differences such as capitalization, number or gender variations.

To report inter-annotator agreement, three annotators ranked a smaller sample of 120 translations: 10 sentences per dataset translated by the four MT engines. The agreement between the three annotators using Cohen's Kappa was 0.572, which according to Landis and Koch (1977) is considered a moderate agreement.

Table 17 shows the ranking of sentences per system, considering the 75 sentences (4 translations for each sentence, 300 translations in total). Online-G clearly contrasts with the other systems, having produced 50 translations that were considered the best and only 6 that were ranked in fourth place. It is important to note that the total of 90 translations that ranked first place (instead of the expected 75) results from ties that occur, for example, when multiple systems produce perfect translations.

System	1	2	3	4
Online-S	11	31	9	24
Online-G	50	14	5	6
Moses-PSMT	17	16	31	11
Moses-HSMT	12	14	21	28
Total	90	75	66	69

Table 17 Number of sentences per ranking level and MT system

With this ranking, we were able to see how many times each system was better than the other (Table 18). These results show that Online-G ranks higher than the other systems the majority of times. Online-S is the next best system, ranking 98 times higher than other systems. Finally, Moses-PSMT and Moses-HSMT are the systems that ranked the lowest (88 and 54 times, respectively).

better \ than	**Online-S**	**Online-G**	**Moses-PSMT**	**Moses-HSMT**	**Total**
Online-S	–	15	41	42	98
Online-G	58	–	56	60	174
Moses-PSMT	34	16	–	38	88
Moses-HSMT	29	11	14	–	54
Total	121	42	111	140	–

Table 18 Comparison between MT systems on 300 translations

6.2 Relating error types with translation quality

Although ranking translations allow us to compare performance between systems, it is not enough to analyze error severity. If we used this ranking in a straightforward manner, we would end up grouping translations that are close to perfect with translations that contain errors that severely hinder comprehension. For instance, the translation *Quem era o Galileo?* (*Who was the Galileo?), which only problem is the insertion of the article "the", ranked the lowest in its group (4^{th} place) because all the translations from the other systems were perfect. This contrasts with the case of a sentence that is placed last because it has severe comprehension problems. Consequently, if used directly, the ranking strategy just described (Section 6.1) places the errors from these two distinct cases at the same level. In other words, it causes a discrepancy of error gravity within the same level.

To overcome this obstacle, we decided to assign a class to each set of 4 translations of the same sentence, using the following criteria:

- **All-Good**: contains sets of translations where all four of them are good, and thus the differences in their rank are caused by minor errors;
- **All-Bad**: is composed of sets of four translations where each has significant problems and errors that hinder comprehension;
- **Mixed**: contains the remaining cases, i.e., sets of translations where some of them have only minor errors (or are perfect) while others have severe comprehension problems.

In sum, this formulation assigns a class to each set of 4 translations according to the differences in terms of quality between them. Having done that, we can now look at what happens inside each class individually, avoiding the problems that were pointed out in the aforementioned example. An error responsible for a lower rank in the All-Good class (like the article insertion in *Who was the Galileo?) does not have the same impact as an error that causes a translation to rank lower on the Mixed or All-Bad classes.

It is also important to mention that the decision to use three classes instead of a finer-grained classification is due to the low agreement reported in the similar task of judging fluency and adequacy in a 5-point scale (Callison-Burch et al, 2007). In our task of grouping sets of translations by quality we achieved a *kappa* of 0.677 (considered substantial agreement). The distribution along classes was of only 14 groups of translations assigned the class All-Good, 21 classified as All-Bad, and the majority (40) ending up classified as Mixed.

This latter class, Mixed, is the one that we are most interested to look at in detail. In it, we can truly distinguish between ranks, since translations that ranked higher are expected to be good while translations that were placed last are likely to have severe comprehension problems. For this reason, in Figure 3 we plotted the average percentage of each error type for the translations in the Mixed class as they occur across ranks. Note that the "Level 1" column contains all translations that ranked first in their set, independently of the system that generated them. Levels 2, 3 and 4 represent the second, third and worst translations, respectively.

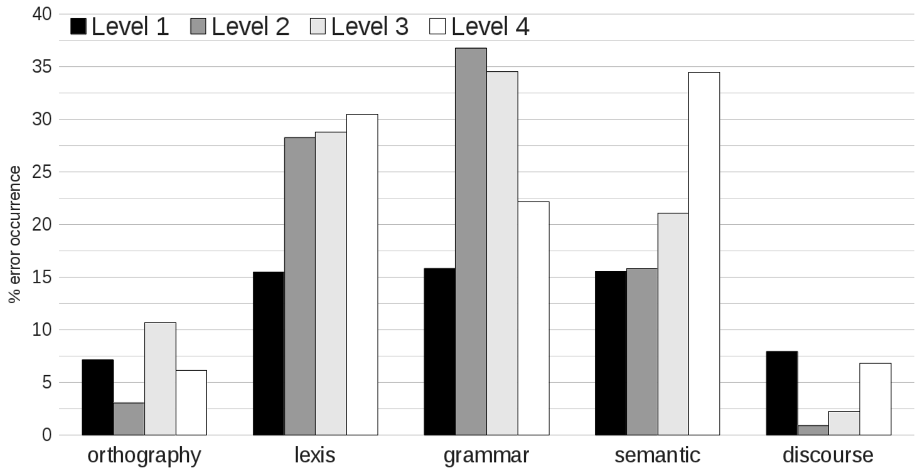

Fig. 3 Percentage of errors on Mixed Class

We can conclude that:

– `lexis` and `semantic` errors are clearly correlated with ranking. Results show that the more `lexis` and `semantic` errors present, the worst the translations rank. This trend is much more accentuated for the `semantic` error type;
– lower percentages of `grammar` errors (approx. 15%) seem to be associated with better quality translations. However, it is not clear how these errors affect translations in lower ranks;
– finally, `orthography` and `discourse` errors have the lowest percentage (< 15% in all ranks), and do not show an increasing or decreasing trend, seeming to occur at the same rate in opposing ranks.

Given some of the inconclusive results obtained in the previous figure, we decided to look further into the taxonomy and plot the subtype of errors and corresponding ranking, again in the Mixed Class (Figure 4). For clarity, we have omitted all errors with low representation (< 1%).

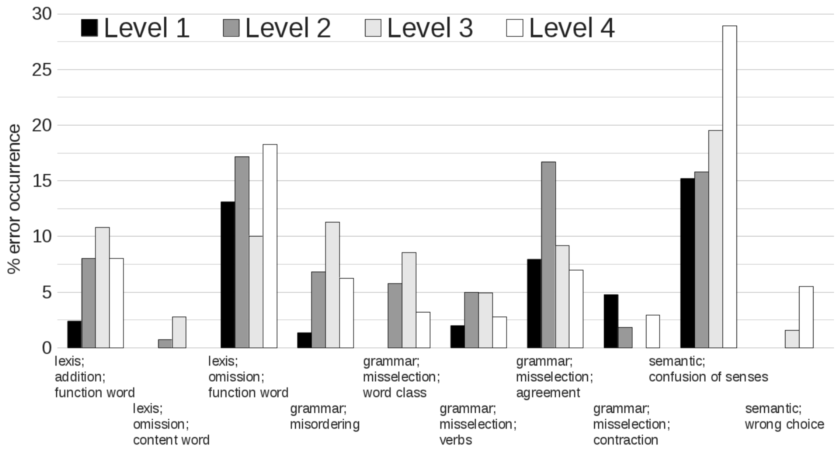

Fig. 4 Percentage of subtype of errors on Mixed Class

When plotting error subtypes, we can see that:

– both subtypes of `semantic` errors show great correlation with ranking. The `wrong choice` subtype only occurs on translations ranked as 3 and 4. The `confusion of senses` subtype demonstrates an exponential growth as the ranking decreases, achievgrammaring almost 30% on translations that ranked last;
– for the remaining error subtypes, both lexical and grammatical, we can not identify a clear trend related to the ranking, with higher percentages of errors being assigned to the central ranks.

Finally, we decided to look into what happens when all translations have severe problems. Figure 5 shows the average percentage of errors for translations in the class All-Bad. Again, we omit error types with low percentages (< 1%).

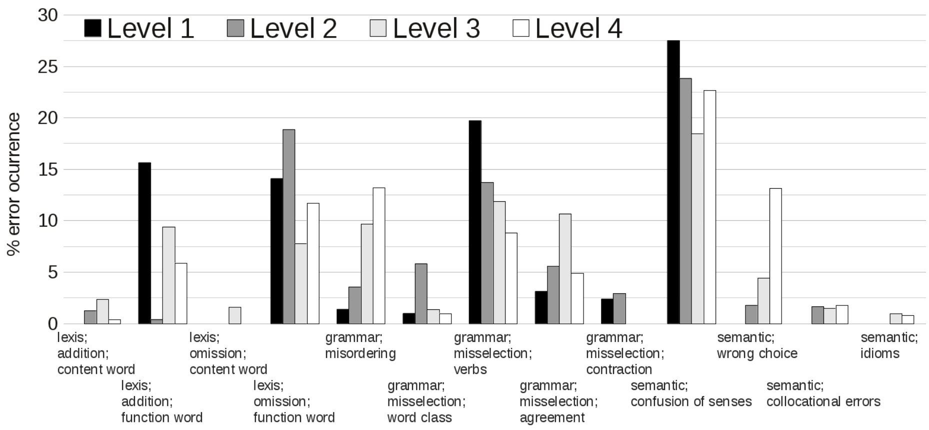

Fig. 5 Percentage of subtype of errors on All-Bad class

We can see that:

- contrarily to what happened to the Mixed class, `confusion of senses` is not discriminative any more. Instead, it just shows a high percentage for all ranks;
- `wrong choice` now shows a more clear correlation;
- also opposed to what happens in the Mixed class, `misordering grammar` errors are now correlated with rank, negatively impacting comprehension;
- other difference is the occurrence of `addition` of content words, `collocational errors`, and `idioms`, absent from the Mixed class, and that appear always associated with lower ranking translations.

7 Conclusions and Future Work

This work aims at developing a detailed taxonomy of MT errors, which extends previous taxonomies, usually focused on English errors. Therefore, the proposed taxonomy is tailored to support errors that are usually associated with morphological richer languages, such as the romance languages. As at the basis of this taxonomy are the main areas of linguistics, we hope that it can be easily extended in order to support errors associated with specific phenomena in other romance languages.

After establishing our taxonomy, we automatically translated three corpora, each one representing a specific translation challenge, by using four different systems: two mainstream online translation systems (Google Translate (Statistical) and Systran (Rule-based)), and two in-house MT systems. Errors were manually annotated, according to the proposed taxonomy, allowing us to evaluate each system and establish some comparisons between them. For instance, we concluded that Online-G has several mistakes of `variety`. Probably, as much of its sources of training are BP and it is not distinguishing between the two varieties, it translates EN into BP and not EP. Regarding Moses-HSMT and Moses-PSMT, we could find many `lexis` errors, as their training corpus is limited in size and in domain. A detailed error analysis is provided in the paper.

Regarding error gravity, we have found that problems related to `confusion of senses`, `wrong choice` and `misordering` are the phenomena that most impacts translation quality, since they seem to correlate with a subjective ranking of the translations.

In what concerns future work, we intend to ask human translators to translate the three corpora to EP. First, we will see if we need to extend our taxonomy in order to support their mistakes. For instance, our taxonomy does not support "invented" words, which are not usual in MT, as most systems only output words that were seen during training. Nevertheless, these errors are usual in human translations. This type of error could be easily integrated in the taxonomy, at the lexis level. Second, we will analyze the attained errors and compare them with the ones committed by translation engines.

Finally, we would like to automatize some steps of our taxonomy. With some statistical learning, errors like, omission, addition and words that were not translated could be automatically found and that would definitely help to make the errors analysis quicker. Also, having information about the most critical sentences to translate could shade some light on where the translation errors might be found.

Acknowledgements

This work was partially supported by national funds through FCT - Fundação para a Ciência e a Tecnologia, under project PEst-OE/EEI/LA0021/2013. Ângela Costa, Wang Ling and Rui Correia are supported by PhD fellowships from FCT (SFRH/BD/85737/2012, SFRH/BD/51157/2010, SFRH/BD/51156/2010).

References

Batista F, Mamede N, Trancoso I (2007) A Lightweight on-the-fly Capitalization System for Automatic Speech Recognition. In: Proceedings of the Recent Advantages in Natural Language Processing (RANLP'07), Borovets, Bulgaria

Bojar O (2011) Analysing Error Types in English-Czech Machine Translation. The Prague Bulletin of Mathematical Linguistics pp 63–76

Bojar O, Mareček D, Novák V, Popel M, Ptáček J, Rouš J, Žabokrtský Z (2009) English-Czech MT in 2008. In: Proceedings of the Fourth Workshop on Statistical Machine Translation, Association for Computational Linguistics, pp 125–129

Callison-Burch C, Fordyce C, Koehn P, Monz C, Schroeder J (2007) (meta-) evaluation of machine translation. In: Proceedings of the Second Workshop on Statistical Machine Translation

Chiang D (2007) Hierarchical Phrase-Based Translation. Comput Linguist 33(2):201–228, DOI 10.1162/coli.2007.33.2.201, URL http://dx.doi.org/10.1162/coli.2007.33.2.201

Condon SL, Parvaz D, Aberdeen JS, Doran C, Freeman A, Awad M (2010) Evaluation of Machine Translation Errors in English and Iraqi Arabic. In: LREC, European Language Resources Association, pp 159 – 168

Corder SP (1967) The Significance of Learner's Errors. International Review of Applied Linguistics pp 161–169

Costa A, Luís T, Ribeiro J, Mendes AC, Coheur L (2012) An English-Portuguese parallel corpus of questions: translation guidelines and application in SMT. In: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), European Language Resources Association (ELRA), Istanbul, Turkey, pp 2172–2176

Costa A, Luís T, Coheur L (2014) Translation errors from English to Portuguese: an annotated corpus. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), European Language Resources Association (ELRA), Reykjavik, Iceland

Cunha C, Cintra L (1987) Nova Gramática do Português Contemporâneo. Edições Sá da Costa, Lisboa

De Saussure F (1916) Cours de linguistique générale. Payot

Denkowski M, Lavie A (2014) Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In: Proceedings of the EACL 2014 Workshop on Statistical Machine Translation

Elliott D, Hartley A, Atwell E (2004) A Fluency Error Categorization Scheme to Guide Automated Machine Translation Evaluation. Lecture Notes in Computer Science, vol 3265, Springer, pp 64–73

Fishel M, Bojar O, Popović M (2012) Terra: a Collection of Translation Error-Annotated Corpora. In: Proceedings of the Eight International Conference on

Language Resources and Evaluation (LREC'12), European Language Resources Association (ELRA), Istanbul, Turkey, pp 7 – 14
H Dulay MB, Krashen SD (1982) Language Two. Newbury House, Rowley
James C (1998) Errors in Language Learning and Use. Exploring Error Analysis. Longaman
Keenan EL, Stabler EP (2010) Language variation and linguistic invariants. Lingua 120(12):2680–2685
Kirchhoff K, Rambow O, Habash N, Diab M (2007) Semi-automatic error analysis for large-scale statistical machine translation. In: In Proc. of the MT Summit XI, Copenhagen, Denmark
Koehn P (2005) Europarl: A Parallel Corpus for Statistical Machine Translation. In: Conference Proceedings: the tenth Machine Translation Summit, AAMT, pp 79–86
Koehn P, Hoang H, Birch A, Callison-Burch C, Federico M, Bertoldi N, Cowan B, Shen W, Moran C, Zens R, Dyer C, Bojar O, Constantin A, Herbst E (2007) Moses: open source toolkit for Statistical Machine Translation. In: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '07, pp 177–180, URL http://dl.acm.org/citation.cfm?id=1557769.1557821
Landis RJ, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33:159–74
Li X, Roth D (2002) Learning question classifiers. In: Proc. 19th Int. Conf. Computational linguistics, ACL, pp 1–7
Ling W, Luis T, Graca J, Coheur L, Trancoso I (2010) Towards a General and Extensible Phrase-Extraction Algorithm. In: IWSLT '10: International Workshop on Spoken Language Translation, Paris, France, pp 313–320
Llitjós AF, Carbonell JG, Lavie A (2005) A Framework for Interactive and Automatic Refinement of Transfer-Based Machine Translation. In: In Proceedings of EAMT 10th Annual Conference, pp 30–31
Niessen S, Och FJ, Leusch G, Ney H (2000) An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. In: LREC, European Language Resources Association
Och FJ (2003) Minimum Error Rate Training in Statistical Machine Translation. In: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1, Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '03, pp 160–167, DOI 10.3115/1075096.1075117, URL http://dx.doi.org/10.3115/1075096.1075117
Och FJ, Ney H (2003) A systematic comparison of various statistical alignment models. Comput Linguist 29:19–51
Papineni K, Roukos S, Ward T, Zhu WJ (2002) BLEU: a method for automatic evaluation of Machine Translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '02, pp 311–318, DOI 10.3115/1073083.1073135, URL http://dx.doi.org/10.3115/1073083.1073135
Popović M, Ney H (2006) Error Analysis of Verb Inflections in Spanish Translation Output. In: TC-STAR Workshop on Speech-to-Speech Translation, Barcelona, Spain, pp 99 –103

Popović M, Gispert Ad, Gupta D, Lambert P, Ney H, Marino JB, Federico M, Banchs R (2006) Morpho-syntactic Information for Automatic Error Analysis of Statistical Machine Translation Output. In: Human Language Technology Conf/North American Chapter of the Assoc. for Computational Linguistics Annual Meeting (HLT-NAACL), Workshop on Statistical Machine Translation, Association for Computational Linguistics, New York City, pp 1–6

Popović N Maja, Hermann (2011) Towards automatic error analysis of Machine Translation output. Comput Linguist 37(4):657–688

Richards J (1974) Error Analysis. Perspectives on Second Language Acquisition. Longaman

Secară A (2005) Translation Evaluation – a State of the Art Survey. In: eCoLoRe/MeLLANGE Workshop, Leeds, UK, pp 39–44

Vilar D, Xu J, D'Haro LF, Ney H (2006) Error Analysis of Machine Translation Output. In: International Conference on Language Resources and Evaluation, Genoa, Italy, pp 697–702

Zeman D, Fishel M, Berka J, Ondřej (2011) Addicter: What Is Wrong with My Translations? Prague Bulletin of Mathematical Linguistics 96:79 – 88