# A Comparison of MT Errors and ESL Errors

**Homa B. Hashemi, Rebecca Hwa**

Intelligent Systems Program, Computer Science Department
University of Pittsburgh
Pittsburgh, PA 15260 USA
hashemi@cs.pitt.edu, hwa@cs.pitt.edu

**Abstract**

Generating fluent and grammatical sentences is a major goal for both Machine Translation (MT) and second-language Grammar Error Correction (GEC), but there have not been a lot of cross-fertilization between the two research communities. Arguably, an automatic translate-to-English system might be seen as an English as a Second Language (ESL) writer whose native language is the source language. This paper investigates whether research findings from the GEC community may help with characterizing MT error analysis. We describe a method for the automatic classification of MT errors according to English as a Second Language (ESL) error categories and conduct a large comparison experiment that includes both high-performing and low-performing translate-to-English MT systems for several source languages. Comparing the distribution of MT error types for all the systems suggests that MT systems have fairly similar distributions regardless of their source languages, and the high-performing MT systems have error distributions that are more similar to those of the low-performing MT systems than to those of ESL learners with the same L1.

**Keywords:** Error Analysis, Machine Translation, English as a Second Language

## 1. Introduction

Performing error analysis of machine translation output is an important but challenging task. Because there is no unique "gold standard" translation for any text, it is difficult to define in-depth evaluation measures. Existing automatic metrics such as BLEU (Papineni et al., 2002) and Meteor (Denkowski and Lavie, 2011) primarily provide a single-value evaluation of the quality of the translation but the task of improving a translation system needs more detailed information about identifying source of errors in a given system. On the other end of the spectrum, manual evaluation provides most reliable error analysis, but it is time-consuming and costly. Even in this case, we still need to come up with standards and guidelines for characterizing the error types.

This paper investigates whether research findings from the Grammar Error Correction (GEC) community may help with characterizing MT error analysis. GEC is concerned with the automatic detection of common grammar mistakes made by students while learning a language that is not their native tongue (e.g., English as a Second Language (ESL)). Arguably, an automatic translate-to-English system might be seen as a non-native English writer whose native language is the source language. The intuition is that since MT systems and non-native speakers of English both do not have a perfect model of fluent English, they may carry over some linguistic properties of the source language that do not hold in English.

Previous work have shown that learners with the same native language (i.e., the same L1) tend to make similar types of mistakes when they learn English (Wong and Dras, 2009; Leacock et al., 2010; Wong and Dras, 2011; Dahlmeier and Ng, 2011; Rozovskaya and Roth, 2011). One of the goals of this paper is to determine whether MT systems of a source language make mistakes in ways that are similar to ESL learners of the same L1. If this is true, then MT models might be improved by addressing those regular, predictable error patterns for a given source language. A second goal of this paper is to determine the frequency with which MT errors fall into one of the ESL categories. If certain ESL error types are common for MT errors, then GEC methods might be applied to correct them.

In this paper, we first describe a method for the automatic classification of MT errors according to ESL error categories developed by Rozovskaya and Roth (2010). We validate the performance of the automatic method empirically. Next, we conduct a large comparison experiment that includes both high-performing and low-performing translate-to-English MT systems for several source languages. Comparing the distribution of MT error types for all the systems, we find that, unlike human ESL learners, the MT systems in our experiment do not seem to be very sensitive to the source languages; their error distribution patterns are relatively similar to each other and not very similar to the ESL learners. With respect to the second goal, we observe that a non-negligible portion of the MT errors do fall into some ESL error categories, even though translation word choice errors are still the majority.

## 2. Automatic MT Error Analysis

In an earlier work, Rozovskaya and Roth (2010) compared common grammar errors made by ESL students with different L1. In this paper, we follow their categories, described in Table 1. However, while the ESL student writing sample is small enough for manual annotation, we want to perform MT error analysis on a large sample of translation outputs, so manual analysis is not practical. Therefore, we need to develop a method to automatically identify and extract translation mistakes.

Figure 1 shows the general procedure for our proposed method of automatic error analysis. The basic idea is to align hypothesis sentence with its corresponding reference

| Error type | Description | Example |
|---|---|---|
| **Article error** | Errors involving an article | "...to focus on [the/None] football as a whole." |
| **Preposition error** | Errors involving a preposition | "...[to/for] outpatient treatment." |
| **Noun number** | Confusion between plural/singular noun | "...where [the families/family] and friends ..." |
| **Verb form** | Verb tense or verb inflection errors | "...the European commission [warns/having warned] against ..." |
| **Word form** | Same stem with wrong suffix | "...where [the tax/taxation] is particularly favorable." |
| **Word choice: insertion, deletion or replacement** | Other errors that cannot be categorized into the above categories | "He also promised to [resolve/solve] the bohemians case." |

Table 1: ESL error types used in Rozovskaya and Roth (2010) for annotation and in this study for automatic error classification. The examples are the MT output sentences and the parts that are different from reference sentences are located inside brackets. The delimiters separate the MT system and reference sentence choice of words.

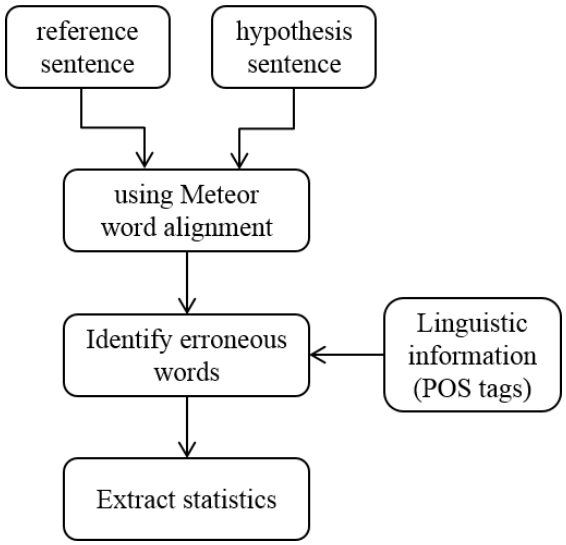


Figure 1: General procedure for automatic error analysis based on the Meteor word alignment and POS tags.

to identify all erroneous words and then use their part-of-speech (POS) taggings to categorize errors.

For establishing the alignments between hypotheses and reference sentences, we use Meteor (Denkowski and Lavie, 2011) word alignment, which has a multistage process based on exact, stem, lexical synonym, and paraphrase matches between words and phrases.

Based on the alignments, we classify different types of translation errors using the following procedure:

- Unaligned reference words are marked as missing. They are further classified into missing article, preposition, punctuation and content word using POS tags.
- Unaligned hypothesis words are marked as inserted words and classified based on their POS tags into extra article, preposition, punctuation and content word.
- Aligned words with different surface form and POS tags are marked as word form, verb form, noun number and word replacement errors.

In order to detect article confusion and preposition replacement errors, when two words are exactly aligned, their previous words are compared, if they were two different articles or prepositions, they would be marked as article confusion or proposition replacement errors. Also, if they were marked previously, the old marks would be deleted. An example of a reference sentence and hypothesis sentence along with the corresponding word alignments is shown in Figure 2.

## 3. Experiments

We conduct two experiments. In the first experiment, we evaluate the quality of the proposed automatic error extraction method by comparing its outputs against a small set of manually constructed gold standard. In the second experiment, we apply the automatic error extractor over a large sample of MT outputs to build distributions of error types for MT systems under different conditions. We compare these error distributions with the error distribution patterns of ESL learners reported by Rozovskaya and Roth (2010).

### 3.1. Data

Our experiments are conducted using the reference sentences and system output submissions to WMT12[1] shared tasks. The WMT12 reference data consists of 3003 sentences which are used in our automatic error analysis approach. To control for the variation of source languages and the quality of the MT systems, our experiment compares a total of eight systems: a high-performance and a low-performance system for each of the following language pairs: German-to-English, Czech-to-English, Spanish-to-English and French-to-English (Callison-Burch et al., 2012).

### 3.2. Verification of Automatic Error Extraction

We manually annotated a subset of the translation outputs so that the automatically flagged errors can be verified. In this set of manual annotation, we limit the scope of project and annotate mistakes of the reported highest-performing and lowest-performing MT systems of WMT12 German-English shared task. We randomly picked 100 reference sentences and then select their corresponding translation outputs from high and low-performing systems. So, we manually annotated totally 200 sentences according to the ESL taxonomy (see Table 1) and compare the performance of our method in the form of precision and recall based on the number of correct detected errors.

Table 2 presents the individual precision, recall and F-scores for each error type. It can be seen that in comparison to human annotator, the automated method has an acceptable precision and recall rate for most categories. The

[1] http://www.statmt.org/wmt12/

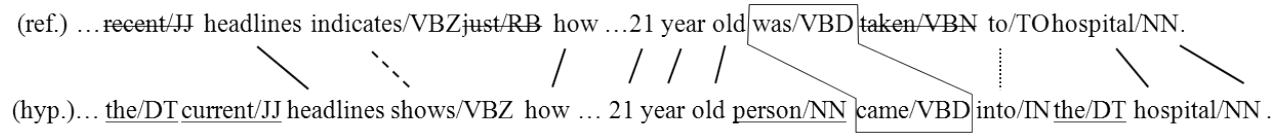


Figure 2: Meteor's (v. 1.4) multi-stage alignment using exact match (solid line), synonym match (dashed line), word form match (dotted line) and paraphrase (solid outline). Missing words in reference sentence (strikethrough) and inserted words at hypothesis (underlined).

| | high-performing MT | | | | low-performing MT | | | |
|---|---|---|---|---|---|---|---|---|
| **Error types** | **# of Manual errors** | **P** | **R** | **F** | **# of Manual errors** | **P** | **R** | **F** |
| All errors | 1455 | 0.76 | 0.81 | 0.79 | 2142 | 0.85 | 0.86 | 0.85 |
| Article error | 136 | 0.87 | 0.63 | 0.73 | 183 | 0.88 | 0.81 | 0.84 |
| Preposition error | 183 | 0.72 | 0.74 | 0.73 | 284 | 0.78 | 0.75 | 0.77 |
| Verb form | 38 | 0.55 | 0.55 | 0.55 | 30 | 0.26 | 0.37 | 0.31 |
| Word form | 29 | 1 | 0.34 | 0.51 | 23 | 1 | 0.3 | 0.47 |
| Noun number | 8 | 0.83 | 0.63 | 0.71 | 22 | 1 | 0.55 | 0.71 |
| Word choice | 947 | 0.74 | 0.86 | 0.8 | 1411 | 0.86 | 0.89 | 0.87 |
| Punctuation | 114 | 0.92 | 0.96 | 0.94 | 189 | 0.94 | 1 | 0.97 |

Table 2: Evaluation results of our automatic error analysis method: precision($P$), recall($R$) and F-score($F$) of each ESL error type based on the manual annotated errors.

overall F-scores range between 79-85%. Although the automatic extraction method is not perfect, we believe that it is sufficient for performing comparative error analysis. Moreover, the automatic method allows us to collect many more instances than manual error analysis so that trends may be observed on a larger scale.

### 3.3. Automatically Extracted Errors

In this experiment, we have applied the automatic error analysis approach described in Section 2. over the outputs of a high-performance and a low-performance MT system for four language pairs. Table 3 summarizes the results. As a point of comparison, we also display the error distributions of ESL learners as reported by Rozovskaya and Roth (2010).[2] A limitation of this comparison study is that the corpus of corrected ESL writings is quite different from the MT dataset; It is a short collection of 200-300 sentences of student essays (per L1). The ESL student mistakes are manually corrected, and in some cases, multiple alternatives are given; in contrast, the MT errors are automatically detected, so mismatches with the references are typically considered an error. Despite these data differences, by examining the distributions of the error types, we might make some qualitative comparisons. To calibrate the quality of each system's performance across different corpora, we report the the number of errors made per 100 words.

The objectives of this experiment are to determine to what extent is there a relationship between MT systems and ESL learners. Our two hypotheses are:

- The types of errors made by MT systems are dependent on the source language, just as ESL learners' errors are impacted by their L1.
- The distribution of error types for high-performing MT systems may be more similar to that of the ESL students than to low-performing MT systems.

Comparing the error distributions of the eight MT systems, we observe that all eight MT systems have fairly similar distributions regardless of their source languages. Moreover, the four high-performing MT systems have error distribution that are more similar to those of the low-performing MT systems than to those of ESL learners with the same L1. Thus, the results suggest that our two hypotheses do not hold. However, it is also not the case that the source language has no impact on the performance of the systems. For example, consider the German-English case. Comparing to other ESL students, German students make a significant amount of punctuation mistakes (51%). In a more subdued fashion, the low-performing German-English MT system makes more punctuation errors than the high-performing system. In contrast, for the other languages the low and high-performing machine translation systems make similar proportions of punctuation mistakes. There are two possible explanations for why the error distribution patterns of MT systems behaved differently than what we hypothesized. One is that the MT systems have largely the same underlying model – they are statistical phrase-based systems. They use the same English language model; therefore, they have similar preferences in terms of the fluency of the output sentences. Another possibility is that because MT systems are still struggling with word choice decisions, which is an adequacy problem, their distribution of errors is skewed by the high *word choice* errors, which are mostly due to unaligned words in the reference and MT output sentences. In contrast, the focus of the ESL error categories is on fluency problems only. To factor out the impact of *word choice* error, we examine article errors in greater detail.

#### 3.3.1. Statistics on Article Errors

Article errors are common mistakes for ESL learners (Izumi et al., 2004; Tetreault and Chodorow, 2008; Ro-

[2]For the ESL errors, we excluded *spelling* and *word order* errors and renormalized the distribution with the remaining errors.

| Language | Method | Errors per 100 words | ESL Error Type | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | Articles | Prepo-sitions | Verb form | Word form | Noun number | Word choice | Punct-uation |
| German-English | ESL | 10.5 | 4.3% | 14.1% | 4.7% | 3% | 2.1% | 16.7% | 55.2% |
| | high-performing MT | 66.1 | 7.4% | 13.3% | 2.8% | 0.4% | 0.5% | 68.2% | 7.2% |
| | low-performing MT | 89.4 | 6.7% | 13.4% | 2.1% | 0.3% | 0.5% | 66.8% | 10.2% |
| Czech-English | ESL | 11.4 | 18.4% | 12.2% | 5.9% | 3.8% | 3.1% | 36.8% | 19.8% |
| | high-performing MT | 67.2 | 8.6% | 13.7% | 2.8% | 0.4% | 0.5% | 67.6% | 6.4% |
| | low-performing MT | 76.2 | 8.4% | 13.0% | 2.4% | 0.3% | 0.4% | 68.6% | 6.9% |
| Spanish-English | ESL | 13 | 13.3% | 16.4% | 6.9% | 4.4% | 3% | 43.6% | 12.4% |
| | high-performing MT | 54.9 | 8.1% | 14.4% | 3.4% | 0.4% | 0.4% | 67.1% | 6.1% |
| | low-performing MT | 72.3 | 6.9% | 12.5% | 2.1% | 0.3% | 0.3% | 69.9% | 7.9% |
| French-English | ESL | 5 | 7.7% | 20% | 2.4% | 4.6% | 5.3% | 14.4% | 45.7% |
| | high-performing MT | 59.4 | 7.9% | 13.9% | 3.6% | 0.5% | 0.5% | 67.5% | 6.3% |
| | low-performing MT | 63.7 | 7.8% | 13.3% | 3.3% | 0.4% | 0.4% | 67.4% | 7.2% |

Table 3: Statistics on the annotated ESL speakers essays and two MT systems for each language.

zovskaya and Roth, 2010; Dalgish, 1985; Han et al., 2006) as well as MT systems. In this subsection, we further expand this error type into more exact sub-categories. As before, we compare the error distributions of the MT systems with each other and with ESL learners. Since the distributions for the MT systems are not skewed by the preponderance of *word choice* errors, the patterns ought to be more directly comparable with those of the ESL learners.

We expand the article error category into six major subcategories: missing *the*, inserting an extra *the*, missing *a*, inserting an extra *a*, *confusion*, for using the wrong article, and *other*, for all other rarer error types. The results are summarized in Table 4.[3] We find that the impact of the source language on MT errors is still not as clearly shown as the L1 is for the ESL learners. With the possible exception of the two Czech-English systems and the low-performance Spanish-English system, the article error distributions for the remaining five systems are fairly similar to each other. In the Czech-English case, it is actually the low-performance MT system whose error distribution pattern looks more similar to the Czech-ESL learner.

In this more restrictive comparison, we still do not observe a strong impact of the source language on the errors made by the MT systems. This suggests that the common English language model used by all the systems may be the cause for the similarity.

## 4. Conclusion

We have investigated a method for automatic SMT error analysis in terms of ESL mistake categories. We conducted experiments to compare high-performing and low-performing MT systems for several language pairs. The results suggest that MT systems have fairly similar distributions regardless of their source languages, and the high-performing MT systems have error distributions that are more similar to those of the low-performing MT systems than to those of ESL learners with the same L1. This may be due to the common English language model component that all the systems use. The experiment does find that MT systems make many errors that fall into one of the ESL error categories, even though making good translation word choices remains the bigger challenge.

[3]For the ESL errors, we excluded the *multiple labels* cases, in which the students' choices are not wrong; we then normalized the distribution with the remaining errors.

## Acknowledgments

This work is supported by U.S. National Science Foundation Grant IIS-0745914. We would like to thank Alon Lavie and Chris Dyer for their valuable comments.

## 5. References

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada, June.

Daniel Dahlmeier and Hwee Tou Ng. 2011. Correcting semantic collocation errors with L1-induced paraphrases. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 107–117.

Gerald Dalgish. 1985. Computer-assisted ESL research. *CALICO Journal*, 2(2):32.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems. In *Proceedings of the EMNLP 2011 Workshop on Statistical Machine Translation*.

Na-Rae Han, Martin Chodorow, and Claudia Leacock. 2006. Detecting errors in English article usage by non-native speakers. *Natural Language Engineering*, 12(02):115–129.

Emi Izumi, Kiyotaka Uchimoto, and Hitoshi Isahara. 2004. The NICT JLE corpus: Exploiting the language learner's speech database for research and education. *International Journal of the Computer, the Internet and Management*, 12(2):119–125.

Claudia Leacock, Martin Chodorow, Michael Gamon, and Joel Tetreault. 2010. Automated grammatical error detection for language learners. *Synthesis lectures on human language technologies*, 3(1):1–134.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318.

| Language | Method | Errors per 100 words | **Article mistakes by Error Type** | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | Missing the | Missing a | Extra the | Extra a | Confusion | Other |
| German-English | ESL | 0.4 | 24% | 10% | 24% | 4% | 9% | 29% |
| | high-performing MT | 4.9 | 14.9% | 10.4% | 37.0% | 11.2% | 8.6% | 17.9% |
| | low-performing MT | 6 | 13.8% | 9.7% | 38.8% | 11.2% | 6.1% | 20.4% |
| Czech-English | ESL | 2 | 30% | 14% | 30% | 9% | 7% | 9% |
| | high-performing MT | 5.8 | 16.8% | 13.0% | 34.0% | 10.9% | 11.3% | 14% |
| | low-performing MT | 6.4 | 22.8% | 16.4% | 28.2% | 8.5% | 8.9% | 15% |
| Spanish-English | ESL | 1.7 | 16% | 10% | 53% | 7% | 3% | 10% |
| | high-performing MT | 4.4 | 13.9% | 9.4% | 41.1% | 11.1% | 8.2% | 16.3% |
| | low-performing MT | 5 | 13.7% | 16.4% | 38.4% | 8.8% | 7.1% | 15.6% |
| French-English | ESL | 0.3 | 12% | 18% | 46% | 18% | 0% | 6% |
| | high-performing MT | 4.7 | 17.0% | 10.6% | 35.3% | 11.6% | 8.9% | 16.6% |
| | low-performing MT | 5 | 15.8% | 10.5% | 36.6% | 11.4% | 8.8% | 17% |

Table 4: Distribution of article mistakes by error type, source language and MT system. *Confusion* error type refers to confusing articles *a* and *the*. *Other* denotes confusing articles with demonstrative and possessive pronouns.


Alla Rozovskaya and Dan Roth. 2010. Annotating ESL errors: Challenges and rewards. In *Proceedings of the NAACL HLT Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 28–36.

Alla Rozovskaya and Dan Roth. 2011. Algorithm selection and model adaptation for ESL correction tasks. *Urbana*, 51:61801.

Joel R Tetreault and Martin Chodorow. 2008. Native judgments of non-native usage: Experiments in preposition error detection. In *Proceedings of the Workshop on Human Judgements in Computational Linguistics*, pages 24–32.

Sze-Meng Jojo Wong and Mark Dras. 2009. Contrastive analysis and native language identification. In *Proceedings of the Australasian Language Technology Association Workshop*, pages 53–61.

Sze-Meng Jojo Wong and Mark Dras. 2011. Exploiting parse structures for native language identification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1600–1610.