Assessing the Impact of Speech Recognition Errors on Machine Translation Quality

Nicholas Ruiz nicruiz@fbk.eu
Marcello Federico federico@fbk.eu
Fondazione Bruno Kessler, Trento, 38122, Italy

Abstract

In spoken language translation, it is crucial that an automatic speech recognition (ASR) system produces outputs that can be adequately translated by a statistical machine translation (SMT) system. While word error rate (WER) is the standard metric of ASR quality, the assumption that each ASR error type is weighted equally is violated in a SMT system that relies on structured input. In this paper, we outline a statistical framework for analyzing the impact of specific ASR error types on translation quality in a speech translation pipeline. Our approach is based on linear mixed-effects models, which allow the analysis of ASR errors on a translation quality metric. The mixed-effects models take into account the variability of ASR systems and the difficulty of each speech utterance being translated in a specific experimental setting. We use mixed-effects models to verify that the ASR errors that compose the WER metric do not contribute equally to translation quality and that interactions exist between ASR errors that cumulatively affect a SMT system's ability to translate an utterance. Our experiments are carried out on the English to French language pair using eight ASR systems and seven post-edited machine translation references from the IWSLT 2013 evaluation campaign. We report significant findings that demonstrate differences in the contributions of specific ASR error types toward speech translation quality and suggest further error types that may contribute to translation difficulty.

1 Introduction

Spoken language translation (SLT) systems are composed with an automatic speech recognition (ASR) system that transcribes source language utterances and a machine translation (MT) system that translates the transcripts into a target language. While there is growing interest in constructing tightly-coupled ASR and MT systems that leverage joint training and optimization (He and Deng, 2012, 2013), the dominant approach is to construct a pipeline consisting of a MT system that decodes one or more ASR hypotheses (Ney, 1999; Matusov et al., 2006; Bertoldi et al., 2007; Casacuberta et al., 2008). The individual SLT components are trained and evaluated independently against local optimization metrics that fit each model to its local task, but they do not generalize to overall SLT quality. In particular, the de-facto automatic evaluation metric for speech recognition is Word Error Rate (WER), which classifies ASR errors into three categories corresponding to Levenshtein distance alignments (i.e. insertions, substitutions, and deletions) between a hypothesis and its reference. The linguistic features of the erroneous words and their relative positions in an utterance are not taken into account.

In this paper, we investigate the impact of ASR errors on speech translation quality. In particular, does each type of ASR error contribute equally to the performance degradation of MT outputs, or do specific classes of ASR errors more greatly inhibit the capacity of the translation model and language model to provide adequate translations? Using the results of the International Workshop on Spoken Language Translation (IWSLT)'s ASR and MT tracks in 2013 (Cettolo et al., 2013), we analyze the impact of ASR errors on the translation quality of TED talks when translating with a standard phrase-based statistical machine translation (SMT) system trained on talk transcripts. We measure the increase of translation errors due to ASR errors over the errors associated with translating well-formed speech transcripts. We further analyze the impacts of ASR errors by performing analyses with linear mixed-effects regression models (Searle, 1973): a generalization of linear regression models suited to model responses with fixed and random effects. ASR errors are categorized based on their Levenshtein alignments to the reference transcript. Experiments are performed on data covering eight ASR systems and 580 utterances in the English to French translation direction. We find that certain types of ASR errors inhibit a translation model's ability to model and accurately translate longer phrases more than others, resulting in disjoint translation hypotheses that are difficult to score by the language model.

In Section 2, we describe related work on ASR and MT error analysis. In Section 3, we describe our experimental setup and outline the research questions used to test for differences between the effects of specific ASR error types on translation errors. In Section 4, we measure the correlation between ASR and MT errors; in Section 5 we test the assumption that translation quality is dependent on the word error rate of ASR hypotheses in the SLT pipeline. In Section 6, we analyze the effects of insertion, deletion, and substitution ASR error types on translation quality and test if each error type equally contributes to the increase in translation errors. We confirm our results by testing for interactions between error types, as well as linguistic properties of the ASR errors that may explain an increase in translation errors. Finally, Section 7 provides concluding remarks and suggestions about the utility of our findings.

2 Previous Work

Error analysis has been successfully used in the ASR and MT communities to improve the quality of each task in isolation. One of the pioneering works of error analysis in MT and ASR is that of Vilar et al. (2006), who categorize MT errors into general categories covering missing words, word order, incorrect words, unknown words, and punctuation errors. Each error type is broken down into specific subtypes covering lexical, syntactic, and semantic properties. Certain error types emerge as frequent issues in translation quality, depending on the language pair. Their analysis on the impact of ASR errors on MT shows that over 50% of the MT errors are associated with substitution errors, many of which are morphological errors that otherwise capture the meaning of the source sentence. While they show the distribution of error types in machine translation outputs, they do not elaborate on the impact of each ASR error type on MT metrics.

He et al. (2011) show that the WER score of an ASR output poorly correlates with its BLEU score of the final SLT output. They demonstrate that optimizing ASR feature weights such as the language model and word insertion penalty to minimize WER can lead to suboptimal translations and instead suggest that discriminative training approaches that optimize WER should be replaced with joint ASR-MT log-linear models that directly optimize ASR and MT features on BLEU. They provide examples of ASR errors related to normalization and speaker disfluencies to show that minimizing WER does not necessarily yield optimal translation scores. While the results suggest that WER minimization in ASR is suboptimal in spoken language translation, they do not identify the contribution of the types of ASR errors on translation errors.

In the ASR community, Goldwater et al. (2010) use a statistical analysis framework based on mixed-effects models to analyze the effects of lexical, prosodic, contextual, and disfluency features of individual words on WER in two state-of-the-art ASR systems and show that their effects on ASR quality are dependent on position and context. For example, they show that while disfluencies account for most ASR errors, only fragments, non-final repetitions, and words preceding fragments have a significant impact. We use a similar experimental setup to measure the influence of different ASR error types, expressed as continuous fixed effects, on the increase in machine translation errors over the translations of perfectly recognized utterances.

A number of works have been proposed to mitigate the contextual effects of ASR errors on MT quality by adapting the SMT phrase table. Ananthakrishnan et al. (2013) use *attention-shift decoding* for ASR (Kumaran et al., 2007) to identify reliable *islands* and unreliable *gaps* in an ASR hypothesis. The SMT decoder penalizes phrase translation pairs whose source phrases span across island-gap boundaries. Likewise, the language model penalizes target language n-grams that cover the island-gap boundary in the source phrase. Tsvetkov et al. (2014) augment phrase-based MT translation models with synthetic phrases by identifying word contexts in ASR outputs that contain acoustically confusable phoneme sequences. The source phrase for each bilingual phrase pair is checked for alternative word sequences that have acoustically similar phoneme sequences. Source-side matches are added to the phrase table with the same target language phrase and new phrase table features are added to measure their fluency.

3 Research methodology

Our goal is to analyze the impact of ASR errors on machine translation quality. Using WER as a metric for ASR quality, how do errors in recognizing speech utterances affect the accuracy of a machine translation system that assumes that each source sentence is clean and well-formed?

We perform our experiments on an intersection of the ASR and MT results of the IWSLT 2013 evaluation campaign (Cettolo et al., 2013), which focused on the translation of TED talks. We collect each submitter's English ASR hypotheses on the tst2012 test set and take the subset of the ASR hypotheses that correspond to the reference set for the English-French MT track. A subset of the MT outputs of each system in the MT track was manually post-edited by professionals and served as multiple human references for automatic evaluation. Using these post-edited translations, we construct 3-way data consisting of eight English ASR hypotheses for 580 utterances, a single unpunctuated reference transcript from the ASR track, and the human post-edited translations from the English-French MT track.

We will use Translation Edit Rate (Snover et al., 2006) as a sentence-level MT quality metric, as it models the original post-editing scenario of the evaluation campaign by estimating the amount of effort required to correct machine translation output. In order to analyze the impact of ASR errors on MT quality, we construct experiments to address the following questions:

- Does ASR's WER correlate with SMT's automatic quality metrics (e.g. TER)?
- Do higher WER scores cause a degradation in MT quality with respect to translations on perfectly recognized utterances (ΔTER)?
- Which types of ASR errors have the strongest impact on translation quality?

In Section 3.1, we discuss the preprocessing steps for each ASR hypothesis and in Section 3.2, we discuss how machine translation outputs are generated for each ASR hypothesis.

3.1 ASR data processing

IWSLT's ASR submissions are in lowercase, lack punctuation, and do not have embedded segmentation. We use the segmentation file provided in the SLT track to induce segmentation. After segmentation, we use the documentation provided in the IWSLT evaluation campaign to

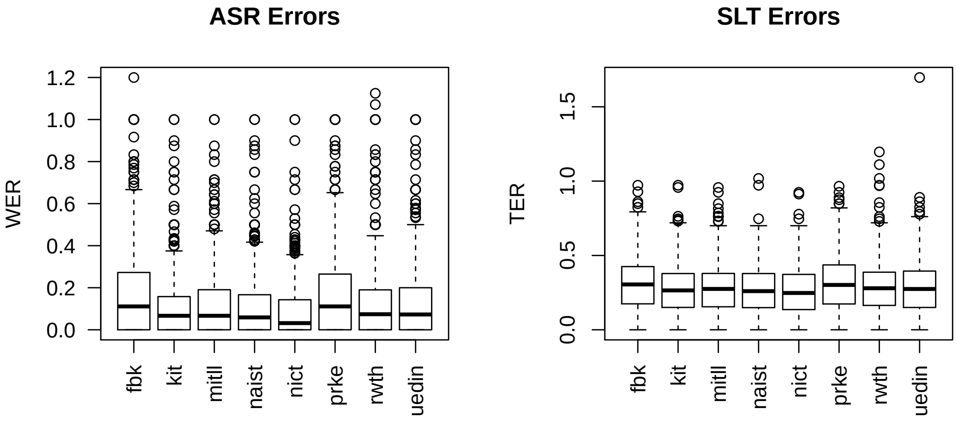


Figure 1: Boxplots describing the distribution of ASR errors (WER) and their impact on translation errors (TER) by ASR system and utterance.

find and match each source transcript and ASR hypothesis with the tst2012 set from the MT track.

Prior to evaluating hypotheses from ASR systems, the DARPA Hub-4 evaluation plan (Pallett et al., 1998) and subsequent ASR evaluations such as NIST's Rich Transcription tasks (Garofolo et al., 2002) used an evolving normalization script to prevent penalization for minor orthographic variations such as multiple spellings (e.g. British vs. American English), compound words (e.g. "storyline" vs. "story line"), and contractions (e.g. "it's" vs. "it is"). Assuming that a phrase-based SMT system in the SLT pipeline is trained on ASR reference transcripts, orthographic variances in ASR outputs can result in out-of-vocabulary words or under-represented source language n-grams in the translation model, further degrading machine translation quality. Although both the ASR hypotheses and the reference transcripts were normalized in prior evaluations, our experiments require the ASR reference to remain unmodified in order to properly evaluate the translation of ASR outputs against translation of the original TED transcripts. Instead, we wrote a supervised word compounding script that splits or compounds words, depending on the word form in the reference transcript. Afterward we apply a bare-bones version of the normalization file provided by IWSLT which only maps British English to American English, since we observed anomalies including inconsistent mappings in the filters used for previous evaluations.

We observed a $\pm 0.3\%$ absolute difference between our WER measures after normalization and the scores reported in the official evaluation task (Cettolo et al., 2013). The rankings of each system remained consistent. In Table 1 we report the performance of each ASR system, before and after orthographic normalization. Note that 5% of the errors for each system are attributed to normalization issues of compounding or word form (e.g. British English instead of American English). The majority of the errors are related to word compounding. The left-hand side of Fig. 1 shows a system-by-system comparison of ASR error distributions. Only a couple of ASR systems have significantly different error distributions from one another.

		ASR WER % ↓				MT TER % ↓	
ASR System	Norm	All	S	D	I	Post-edit	REF
fbk	none	21.4	13.3	2.9	5.2	33.70	54.68
	+COMP	16.8	10.8	3.0	3.0	32.84	54.10
	+NORM	16.5	10.5	3.1	2.9	32.71	54.09
kit	none	15.3	9.2	1.6	4.5	29.86	52.07
	COMP	10.4	6.6	1.7	2.1	28.83	51.40
	+NORM	10.1	6.3	1.7	2.1	28.73	51.40
mitll	none	16.4	9.6	2.0	4.8	30.13	52.17
	COMP	11.6	7.0	2.1	2.5	29.36	51.53
	+NORM	11.4	6.8	2.2	2.4	29.32	51.58
naist	none	15.7	9.1	2.2	4.4	29.86	51.88
	COMP	10.9	6.5	2.3	2.0	28.94	51.31
	+NORM	10.6	6.3	2.3	2.0	28.82	51.28
nict	none	14.5	8.7	1.4	4.4	28.92	51.43
	COMP	9.5	6.0	1.5	2.0	27.94	50.75
	+NORM	9.2	5.8	1.5	1.9	27.84	50.77
prke	none	21.3	13.2	2.8	5.3	33.79	54.83
	COMP	16.9	10.8	2.9	3.1	33.09	54.42
	+NORM	16.6	10.6	2.9	3.1	33.01	54.42
rwth	none	16.5	10.1	1.7	4.7	30.93	52.66
	COMP	11.9	7.7	1.8	2.4	29.93	52.08
	+NORM	11.7	7.5	1.8	2.4	29.84	52.06
uedin	none	17.2	10.2	2.1	4.9	30.84	52.66
	COMP	12.6	7.7	2.3	2.7	29.99	52.04
	+NORM	12.3	7.4	2.2	2.6	29.94	52.05
gold	none	0.0	0.0	0.0	0.0	21.27	46.46

Table 1: ASR outputs used as English-French MT evaluation input data on the human evaluation task of IWSLT 2013. ASR outputs are evaluated with no additional normalization, oracle word compounding (COMP), or compounding with word form normalization (NORM). Translated ASR outputs are tokenized and evaluated against the reference translation (Auto) and a combination of the human post-edited sentences from the MT task (Post-edit).

3.2 MT data processing

Since we are evaluating the impact of ASR errors on translation quality, we use a fixed SMT system trained on TED talk transcripts, which correspond to the reference transcripts in the ASR track, with the addition of punctuation. We use FBK's primary phrase-based SMT system used in the English-French MT track (Bertoldi et al., 2013). The normalized ASR hypotheses are post-processed by inserting punctuation and applying recasing. We insert the punctuation as closely as possible to the position dictated in the reference in order to control the impact of punctuation on translation output. This is done by computing the Levenshtein alignments between the unpunctuated TED transcripts and each ASR hypothesis, using SCLITE[1]. We train and apply a recaser model using the standard Moses tools (Koehn et al., 2007) with IWSLT 2013's TED training data to all of the newly-punctuated ASR outputs.

After introducing punctuation and recasing the ASR output, we translate each ASR output and evaluate the results using TER over the seven human post-edited translations. Translation

[1] http://www1.icsi.berkeley.edu/Speech/docs/sctk-1.2/sclite.htm

results are contrasted with FBK's primary MT submission on the right-hand side of Table 1. We observe over a 4% absolute increase in TER for each of the translations of ASR hypotheses against the reference translation and likewise over a 6% absolute increase against the post-edited references. Most of the ASR transcripts' translations yield a TER score around 30% against the post-edited references. Turchi et al. (2013) empirically determine that translations with a TER score above 40% against a set of post-edited references require the translator to re-translate the source sentence from scratch, while lower scores imply that it is productive for the translator to post-edit the MT output. Likewise, our reported TER scores suggest that the translations of the ASR hypotheses are of good enough quality to be used in a post-editing scenario. The right-hand side of Fig. 1 shows a system-by-system comparison of SLT error distributions, measured in TER. In particular, we observe less variance among ASR systems as their hypotheses are translated by the SMT system.

4 Do ASR errors correlate with SMT errors?

Using the WER metric, how do ASR errors correlate with SMT errors? We split this question into two related questions. First, is there any relation between an ASR system's difficulty to recognize a speech utterance and the difficulty of translating the utterance, assuming it was recognized perfectly? The answer may seem obvious, since an ASR model could be trained poorly and generate hypotheses that have no bearing with their references. However, as described in the previous section, each of the ASR systems used in the IWSLT evaluation are capable of producing translations that can be efficiently post-edited by a professional translator. Second, do ASR errors correspond directly with translation quality? In other words, does the increase or decrease in WER correlate with the number of translation errors in the speech translation pipeline? We address these questions by analyzing the correlation between the independent variable (WER) and the dependent variables (TER on translations of ASR references and ASR hypotheses, respectively) in Section 4.1, followed by constructing linear regression models to test for statistical significance in Section 4.2.

4.1 Correlation

We first measure the correlation between the WER scores of each ASR system and the TER acquired by translating each corresponding ASR reference. The Pearson correlation coefficient, r, measures the linear dependence between two variables. For our experiments, we control the effects of sentence length by binning the ASR hypotheses from each system into buckets corresponding to the quartiles of the reference length. Since much of the skewness of ASR errors shown in Fig. 1 is related to ASR reference length, we take correlation measurements on the 2nd and 3rd length quartiles, corresponding to reference lengths of 9-15 and 16-22. Using all ASR systems, we observe r values of 0.039 and 0.091, on the respective reference lengths, implying no correlation. Using only the observations of NICT's primary system (which had the lowest WER in the ASR evaluation track), we observe r values of -0.031 and 0.049, respectively.

We repeat the experiment, this time comparing ASR errors to their corresponding translation errors. Using all ASR systems, we observe r values of 0.672 and 0.632, respectively, implying strong correlation. We observe a similar trend when considering NICT's system alone. Again, these results are not surprising, since a machine translation system depends on the speech recognition output in order to generate a translation. It is important to note that while there is naturally a strong correlation between ASR outputs and the quality of their translations, translation quality is not solely dependent on ASR quality. The missing 30% includes phenomena related to the problem of transferring content from the source language (English) to the target language (French), which take into consideration the lexical, syntactic, and semantic properties of each language (Vilar et al., 2006; He et al., 2011; Ruiz and Federico, 2014).

4.2 Linear Regression

To verify whether the correlation results in the previous section imply dependence, we fit univariate linear regression models using a single ASR system to evaluate the contribution of WER to the corresponding translation's TER score. We focus our attention on the observations of NICT's primary system. The response variables are the TER scores computed against seven post-edited translation references. TER is computed either on the ASR references or on the translations of NICT's ASR hypotheses. Again, WER is computed on the uncased, unpunctuated output of the ASR system. Translations are performed using FBK's primary MT submission.

Our first experiment estimates the effects of WER on TER acquired by translating each corresponding ASR reference. As suggested by the low Pearson correlation scores in our previous experiment, WER is not a significant predictor of TER scores on the translation of ASR references: $\beta = 0.028, t(578) = 0.719, p = 0.473$, with a negative adjusted r^2 value.

Our second model treats the TER of the translated ASR hypotheses as the response variable. WER significantly predicts TER scores, $\beta = 0.696, t(578) = 18.42, p < 10^{-4}$ and explains a significant proportion of variance in TER scores ($r^2 = 0.369, F(1, 225) = 339.4, p < 10^{-4}$). However, much of the variance remains unexplained by the model. WER normalizes by the reference transcript's utterance length; however, input length is an important factor that affects the search space in SMT decoding. Thus, WER cannot intrinsically anticipate the difficulty of translating the utterance. As evidence, we sample two utterances recognized by NICT's ASR system, both with WER scores of 20% but having a different number of words in the reference (5 and 25, respectively). The TER scores of their translations are 46.7% and 28.4%, respectively. WER also assumes that each error contributes independently towards the error metric and thus does not measure interactions between multiple errors in an utterance. In phrase-based SMT, the position and density of ASR errors can hinder the translation model's ability to select proper target phrases, as well as affect the reordering model's ability to properly arrange the phrases in the target language.

5 Does a higher WER cause an increase in translation errors?

Our previous experiments in Section 4.1 measured the relationship between WER of ASR hypotheses and TER. While WER is a significant predictor of TER in our simple regression model, it fails to capture the variance in TER associated with the innate difficulty of translating the utterance. As shown in the correlation measurements in the first experiment, WER alone cannot provide reliable estimates of the number and types of errors in a perfectly recognized utterance. To control for the difficulty of translating an otherwise perfect speech recognition hypothesis, we use the difference between the TER associated with translating the perfect ASR reference and the TER associated with translating the ASR hypothesis, labeled as ΔTER:

$$\Delta\text{TER} = \text{TER}_{\text{gold}} - \text{TER}_{\text{ASR}}, \tag{1}$$

where TER_{gold} is the TER score for a perfectly recognized utterance, and TER_{ASR} is the TER score on the translation of the ASR hypothesis. By using ΔTER, we assume that TER_{gold} is the upper-bound on translation quality with the given SMT system. In other words, we assume that a SMT system cannot translate transcripts containing errors better than clean transcripts. We check this assumption in our observation data and note 64 violations out of a total of 4,640 observations covering the outputs of the eight ASR systems (1.4% of the time). As a sanity check, we had two native French speakers evaluate the translation quality of several scenarios where $\Delta\text{TER} < -0.1$. In all cases, the native speakers preferred the MT outputs of translated ASR references over the translations of ASR hypotheses. These violations are likely due to the

	All ASR		NICT+FBK	
Fixed effects	β	Std. Error	β	Std. Error
(Intercept)	8.72e-03	3.14e-03 ○	1.01e-02	3.93e-03 ○
WER	6.30e-01	8.55e-03 ●	6.16e-01	1.46e-02 ●
Random effects	Variance	Std. Dev.	Variance	Std. Dev.
UttID (Inter)	4.50e-03	0.067	3.89e-03	0.062
SysID (Inter)	0.000	0.000	2.57e-06	0.002
Residual	3.74e-03	0.061	3.99e-03	0.063

Table 2: Fixed and random effects for the null model, which measures the effect of WER on ΔTER for English-French SLT. The model is constructed with observations from all ASR systems in IWSLT 2013's ASR Track on the left-hand side and only NICT and FBK's ASR systems on the right. Fixed effects coefficients (β) and standard errors are reported. Random intercepts account for variances by utterance (*UttID*) and ASR system (*SysID*). Statistical significance at $p < 10^{-4}$ is marked with ● and $p < 10^{-2}$ is marked with ○.

greedy alignment heuristics used the TER algorithm to accommodate reordering shifts in the Levenshtein alignment (Snover et al., 2006).

We first measure the correlation between WER and ΔTER using Pearson's r. Following the same approach as Section 4.1, we observe strong correlations on the observations with reference lengths in the middle 50% length quartiles: 0.780 and 0.756 using all ASR systems for utterance lengths of 9-15 and 16-22, respectively, and scores of 0.786 and 0.778 using only NICT's ASR system.

We next verify ΔTER's dependence on WER using linear mixed-effets models, which have been effectively used on linguistic data by Baayen et al. (2008). Mixed-effects models allow us to take into consideration random effects caused by an ASR system and the particular features of each ASR utterance. We use the R (R Core Team, 2013) implementation of linear mixed-effects models in the *lme4* library (Bates et al., 2014). As fixed effects, we enter WER into the model, which we label as Model 1. We provide random intercepts for the utterance (labeled as *UttID*) and ASR system (labeled as *SysID*). The models are fit by maximum likelihood. We use repeated observations of 580 speech utterances by eight ASR systems, yielding a total of 4,640 observations. Fixed effect coefficients and random effects variance for Model 1 are reported in Table 2.[2] Both WER and the intercept are observed as statistically significant. The coefficients suggest that if there are no ASR errors, TER will increase by 0.87%. However, for each percentage point of WER, the TER will further increase by roughly $0.63 \times 0.01 = 0.0063$ (0.63%). We observe a r^2 value of 0.840 for the model, 0.154 of which is attributed to the fixed effects.

As a random effect, *SysID* was not significant, as it has a standard deviation near zero. This behavior is also evident in the boxplots of Fig. 1, implying that the differences between the emitted WER scores and translation TER scores for each ASR system are not significantly different from one another. In order to verify that the random intercept associated with the ASR system is indeed insignificant, we repeat the mixed-effects analysis, using two systems with significantly different WER scores; namely NICT and FBK. Statistics on the fixed and random effects are also listed in Table 2. We again observe near-zero variance for *SysID* and do not observe significant differences in the fixed effects coefficients, implying that the *SysID* random effect has no impact on the model.

[2] Note that the WER and TER values in Table 1 are listed as percentages, while our regression models express the values between 0 and 1.

	Model 2 (no interactions)		Model 3 (interactions)	
Fixed effects	β	Std. Error	β	Std. Error
(Intercept)	8.59e-03	3.14e-03 ○	6.30e-03	3.20e-03 $*$
WER.S	6.50e-01	1.19e-02 ●	6.90e-01	1.42e-02 ●
WER.D	6.02e-01	1.84e-02 ●	6.66e-01	2.12e-02 ●
WER.I	6.03e-01	2.26e-02 ●	6.21e-01	3.26e-02 ●
WER.S×WER.D	–	–	-6.82e-01	1.07e-01 ●
Random effects	Variance	Std. Dev.	Variance	Std. Dev.
UttID (Inter)	4.52e-03	0.067	4.55e-03	0.067
SysID (Inter)	0.0000	0.000	0.0000	0.000
Residual	3.73e-03	0.061	3.69e-03	0.061

Table 3: Fixed and random effects for Models 2 and 3, which measure the effect of ASR error types on ΔTER for English-French SLT. Model 3 includes interactions between error types. Fixed effects coefficients (β) and standard errors are reported; statistically insignificant fixed effects are omitted. Random intercepts account for variances by utterance (*UttID*) and ASR system (*SysID*). Statistical significance at $p < 10^{-4}$ is marked with $\bullet$, $p < 10^{-2}$ is marked with $\circ$, and $p < 0.05$ is marked with $*$.

6 Which types of ASR errors have the strongest impact on translation quality?

Now that we have verified that an increase in WER significantly increases TER, are there significant differences between the effects of individual ASR error types on translation quality? We hypothesize that not all ASR errors are treated equally when ASR hypotheses are used in the speech translation pipeline. To demonstrate this, we construct new mixed-effects models which factorize the WER metric into the components used to compute its score. Recall that the WER for an utterance is computed as:

$$WER = \frac{S + D + I}{L}, \tag{2}$$

where S, D, and I are the number of substitutions, deletions, and insertions in the Levensthtein alignment between the hypothesis and the reference, and L is the ASR reference length (in words). According to (2), we factorize WER into three independent variables, corresponding to the number of occurrences of each error type, normalized by the reference length. As random effects, we continue to use the utterance ID and the ASR system ID. Our null hypothesis states that all length-normalized ASR error types (S, D, I) contribute equally to ΔTER, which is the same as our Model 1 specification in Section 5.

6.1 Do Levenshtein error types have differing levels of importance?

In the alternative hypothesis's mixed-effects model, we enter S/L, D/L, and I/L as fixed effects and maintain the same random effects as Model 1. To simplify the notation in our model, which we label Model 2, we refer to the length-normalized error types in shorthand form (e.g. *WER.S*). The coefficients of the fixed effects of the fitted model are shown in Table 3.

All error type coefficients are statistically significant at the $p < 10^{-4}$ level, with $r^2 = 0.840$. We perform a likelihood ratio test between Model 2 and the null model (Model 1), using the `anova()` function in the *lmerTest* R package. We observe marginal significance at the $p < .1$ level ($\chi^2(2) = 5.558, p = 0.062$), suggesting that Model 2 may better describe the relationship between ASR errors and ΔTER; however, a deeper analysis is required.

6.2 Are there interactions between Levenshtein error types?

While the coefficients in Model 2 listed in Table 3 seem to suggest that substitutions have a greater impact on translation quality than deletions or substitutions, the low level of significance may indicate that the model is missing a predictor. The WER metric in (2) posits that the influence of Levenshtein error types on ASR quality are independent from one another. In other words, it assumes that there are no interactions between error types. Our factorization of WER in Model 2 retains this assumption. We now test for interactions between the reference length-normalized error types in Model 2. We construct a new mixed-effects model (Model 3) that contains all pairwise interactions (e.g. *WER.S* × *WER.D*), as well as the interaction triple (*WER.S* × *WER.D* × *WER.I*). The fixed and random effects of the new model are reported on the right-hand side of Table 3. The *WER.S* × *WER.D* interaction is reported as significant ($p < 10^{-4}$), while the other interactions are statistically insignificant.

We perform likelihood ratio tests between Model 3 and our previous models to verify the significance of the interaction term, again using the `anova()` function in the *lmerTest* R package. We observe a significant difference between Models 1 and 3 ($\chi^2(7) = 47.109, p < 1.78 \times 10^{-8}$), as well as between Models 2 and 3 ($\chi^2(4) = 41.55, p < 2.07 \times 10^{-8}$), confirming the presence of interactions between ASR error types. The impact of substitutions in Model 3 is dependent on the number of deletions that co-occur within a sentence. For example, a sentence with a WER of 10%, solely caused by substitution errors, corresponds to approximately a 7.5% increase in TER ($\beta_0 + \beta_S \times 0.1$). However, a sentence with a WER of 15% with 10% as substitution errors and 5% as deletion errors would expect an increase in TER by $\beta_0 + \beta_S \times 0.1 + \beta_D \times 0.05 - \beta_{SD} \times 0.1 \times 0.05 \approx 0.105$ (10.5%); the interaction term reduces the increase in TER by 0.34%. Thus, we can conclude that not only do Levenshtein alignment-based ASR error types have differing levels of importance, but there also exists an interaction between substitution and deletion errors that reduces their individual impact on translation quality (in terms of ΔTER).

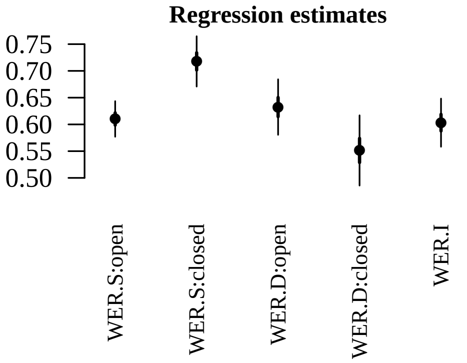
6.3 Are there linguistic patterns of ASR errors that impact translation quality?

In Section 6.1, we tested the hypothesis that individual Levenshtein error types have different effects on translation quality. We showed weakly significant results, suggesting that the breakdown of WER into length-normalized Levenshtein alignment types may better model the relationship between ASR errors and translation quality (in ΔTER). As we have seen with interactions between substitutions and deletions in Section 6.2, there are contexts in which the impact of substitution errors on translation quality may vary. In particular, we believe that linguistically informed errors may better describe how ASR errors violate the structural assumptions used to train standard statistical machine translation systems.

As a preliminary experiment, we focus our attention on misrecognized function words and content words. In particular, researchers such as Goldwater et al. (2010) identify function words (also known as closed class words) as problematic for speech recognition. Oftentimes a speaker may alter the pronunciation of high frequency function words, such as prepositions and articles, by under-articulating or dropping phonemes. While a human can predict these words with high accuracy, an ASR system relies on phoneme or triphone recognition as an intermediate step toward recognizing words. Content words (also known as open class words) are generally simpler to recognize, as they often contain more syllables and cover a larger amount of speaking time within an utterance. On the other hand, open class words might not be represented in a speech lexicon, rendering them impossible to be generated by an ASR system. Aside from the issue of out-of-vocabulary words, SMT systems have the opposite problem. Vilar et al. (2006) demonstrate that missing content words contribute more toward translation errors than missing function words.

	Model 4 (word class)	
Fixed effects	β	Std. Error
(Intercept)	8.95e-03	3.14e-03 ○
WER.S:open	6.10e-01	1.70e-02 ●
WER.S:closed	7.18e-01	2.40e-02 ●
WER.D:open	6.32e-01	2.65e-02 ●
WER.D:closed	5.51e-01	3.35e-02 ●
WER.I	6.03e-01	2.23e-02 ●
Random effects	Variance	Std. Dev.
UttID (Inter)	4.45e-03	0.067
SysID (Inter)	0.0000	0.000
Residual	3.72e-03	0.061

(a) Fixed and random effects for Model 4.



(b) 95% confidence intervals for fixed effects coefficients.

Table 4: Model 4, which measures the effect of open and closed-class ASR error types on ΔTER for English-French SLT. Coefficients (β) with confidence intervals are reported for the fixed effects. Random intercepts account for variances by utterance (*UttID*) and ASR system (*SysID*). Statistical significance at $p < 10^{-4}$ is marked with ● and $p < 10^{-2}$ is marked with ○.

Since we have already observed differences between Levenshtein error types, we now look at differences between how misrecognitions of open and closed class words affect translation outputs. We use TreeTagger (Schmid, 1994) to assign part-of-speech (POS) tags on the ASR references using the Penn Treebank (Marcus et al., 1993). Using the Levenshtein alignments between each ASR hypothesis and its reference, we annotate deletion and substitution errors with their POS tags. We do not annotate insertion errors, as an insertion error indicates that no reference word is available to tag. We manually map each POS tag associated with a substitution and deletion error to its class (open or closed).

Our new model (Model 4) extends Model 2 from Section 6.1 by separating substitution and deletion errors by their word classes. To simplify our model, we do not consider interactions between the error types. Statistics on the fixed and random effects are shown in Table 4a. Our results confirm that all word class-specific ASR error types are significant at the $p < 10^{-4}$ level. Likelihood ratio tests between Models 2 and 4 indicate that the Levenshtein error types grouped by word class better measure the impact of ASR errors on translation quality ($\chi^2(2) = 15.487, p = 4.34 \times 10^{-4}$). The fixed effect coefficients' confidence intervals in Table 4b show that substitution errors on function words have the greatest individual impact on translation errors: every one percent of of these errors increases ΔTER by 0.7% over the intercept, assuming all other factors are held constant. Substitution errors on content words, however, have a significantly lower impact. Conversely, deletion errors on content words have a greater impact than those on function words. All other factors held constant, a standard phrase-based machine translation system is apparently more tolerant of ASR deletion errors on function words than towards substitution errors on function words. We hypothesize that this is most commonly due to cases where a function word is recognized as another function word from a different lexical category (e.g. a preposition recognized as a determiner).

7 Conclusion

In this paper, we focused on the contribution of Levenshtein alignment errors in ASR's word error rate (WER) metric on translation quality in terms of Translation Edit Rate (TER). Working on the English-French translation direction in the IWSLT 2013 TED Talk spoken language translation task, we collected a subset of ASR hypotheses from eight systems on the 2012 test set and measured their translation quality against seven human post-edited translations. Using this data, we measured the impact of ASR errors on TER against a gold standard that measures the inherent complexity of an utterance, assuming perfect speech recognition. We measured the correlation between the WER of ASR hypotheses and the TER of the associated translations, showing that while WER strongly correlates with machine translation evaluation metrics such as TER, it does not account for the inherent complexity of a source language utterance.

We additionally constructed linear mixed-effects models to show that substitution, insertion, and deletion error types in the WER metric do not contribute uniformly to translation errors when using a statistical machine translation (SMT) system trained on clean transcripts. Our results suggest similarly to Vilar et al. (2006) and He et al. (2011) that while WER is the de-facto metric for ASR quality, WER alone is not a good indicator of translation quality due to its assumption of independent error types acting in isolation. WER fails to take into account the cumulative effects of errors, which include interactions between substitution and deletion errors.

Additionally, we provided a preliminary experiment that shows that the linguistic properties of ASR errors have ramifications on SMT quality in speech translation. We annotated substitution and deletion errors by their word class (open or closed), based on the part-of-speech tags assigned to each ASR reference transcript. Our preliminary results show differences in how speech recognition errors on open class and closed class words affect the machine translation engine, particularly due to their Levenshtein alignment types. We see that substitution errors on function words are more detrimental than deleting a function word; though this behavior was not observed for content words. The results indicate that more investigation should be done on the impact of ASR errors by lexical class on translation quality. In our next steps, we plan to train and apply a part-of-speech tagger directly on the ASR output to get a better idea of which types of part-of-speech errors are less likely to be tolerated by a standard phrase-based SMT system.

Our results suggest that additional error types should be considered when measuring the impact of ASR errors on spoken language translation quality. Thus far, our experiments have focused on the role of individual Levenshtein error types and their properties, though, as we have observed with interactions between error types, the context of each ASR error affects how it will impair the translation model and language model from generating a high quality translation. We suggest, for example, to analyze features that account for the distributional density of ASR errors, which may better describe how ASR errors violate the structural assumptions used to train standard SMT systems. For example, let us assume that we have two utterances with the same reference length and WER. If the errors in one utterance are concentrated at the beginning or end of the utterance, would its TER be greater than an utterance whose errors are uniformly distributed across the entire segment?

Finally, while our experiments have focused on MT references generated by human post-editions, we propose to perform this analysis on automatic references with metrics such as NIST (Doddington, 2002), METEOR (Banerjee and Lavie, 2005), and HMEANT (Lo and Wu, 2011). Ultimately, we believe that the analysis of ASR errors on SLT can result in deriving an error metric that better correlates with MT quality in the speech translation pipeline.

Acknowledgments

This work was partially supported by the MateCAT (grant agreement 287688) and EU-BRIDGE projects (grant agreement 287658), which are funded by the EC under the 7th Framework Programme. The authors also thank Matteo Negri and Marco Turchi for their useful suggestions.

References

Ananthakrishnan, S., Chen, W., Kumar, R., and Mehay, D. (2013). Source-Error Aware Phrase-Based Decoding for Robust Conversational Spoken Language Translation. In *International Workshop on Spoken Language Translation (IWSLT)*, Heidelberg, Germany.

Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, 59(4):390–412.

Banerjee, S. and Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2014). *lme4: Linear mixed-effects models using Eigen and S4*. R package version 1.1-6.

Bertoldi, N., Farajian, M., Mathur, P., Ruiz, N., and Federico, M. (2013). FBKs machine translation systems for the IWSLT 2013 evaluation campaign. In *Proc. of the International Workshop on Spoken Language Translation*.

Bertoldi, N., Zens, R., and Federico, M. (2007). Speech translation by confusion network decoding. In *Proceedings of ICASSP*, pages 1297–1300, Honolulu, HA.

Casacuberta, F., Federico, M., Ney, H., and Vidal, E. (2008). Recent efforts in spoken language processing. *IEEE Signal Processing Magazine*, 25(3):80–88.

Cettolo, M., Niehues, J., Stüker, S., Bentivogli, L., and Federico, M. (2013). Report on the 10th IWSLT Evaluation Campaign. In *Proc. of the International Workshop on Spoken Language Translation*.

Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proc. of HLT*, pages 138–145, San Diego, CA.

Garofolo, J. S., Fiscus, J. G., Martin, A. F., Pallett, D. S., and Przybocki, M. A. (2002). Nist rich transcription 2002 evaluation: A preview. In *LREC*. European Language Resources Association.

Goldwater, S., Jurafsky, D., and Manning, C. D. (2010). Which words are hard to recognize? prosodic, lexical, and disfluency factors that increase speech recognition error rates. *Speech Communication*, 52(3):181–200.

He, X. and Deng, L. (2012). Maximum Expected BLEU Training of Phrase and Lexicon Translation Models. *Proceedings of ACL*.

He, X. and Deng, L. (2013). Speech-Centric Information Processing: An Optimization-Oriented Approach. *Proceedings of the IEEE*.

He, X., Deng, L., and Acero, A. (2011). Why word error rate is not a good metric for speech recognizer training for the speech translation task? In *Proc. ICASSP*. IEEE.

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., and Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic.

Kumaran, R., Bilmes, J., and Kirchhoff, K. (2007). Attention shift decoding for conversational speech recognition. In *INTERSPEECH*, pages 1493–1496. ISCA.

Lo, C.-k. and Wu, D. (2011). Structured vs. Flat Semantic Role Representations for Machine Translation Evaluation. In *Proceedings of the Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation*, SSST-5, pages 10–20, Stroudsburg, PA, USA. Association for Computational Linguistics.

Marcus, M., Santorini, B., and Marcinkiewicz, M. (1993). Building a large annotated corpus of english: the penn treebank. *Computational Linguistics*, 19:313–330.

Matusov, E., Kanthak, S., and Ney, H. (2006). Integrating speech recognition and machine translation: Where do we stand? In *Proceedings of ICASSP*, pages 1217–1220, Toulouse, France.

Ney, H. (1999). Speech translation: coupling of recognition and translation. In *Proceedings of ICASSP*, Phoenix, Arizona.

Pallett, D., Fiscus, J. G., Garofolo, J. S., Martin, A., and Przybocki, M. (1998). The 1998 Hub-4 Evaluation Plan for Recognition of Broadcast News, in English. `http://www.itl.nist.gov/iad/mig/tests/bnr/1998/hub4e_98_spec.html`. Accessed: 2014-05-21.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ruiz, N. and Federico, M. (2014). Complexity of Spoken Versus Written Language for Machine Translation. In *Proceedings of the 17th Conference of the European Association for Machine Translation (EAMT)*, pages 173–180.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*.

Searle, S. R. (1973). Prediction, mixed models, and variance components. *Biometrics Unit*, (BU-468-M).

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *5th Conference of the Association for Machine Translation in the Americas (AMTA)*, Boston, Massachusetts.

Tsvetkov, Y., Metze, F., and Dyer, C. (2014). Augmenting Translation Models with Simulated Acoustic Confusions for Improved Spoken Language Translation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 616–625, Gothenburg, Sweden. Association for Computational Linguistics.

Turchi, M., Negri, M., and Federico, M. (2013). Coping with the subjectivity of human judgements in MT quality estimation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 240–251, Sofia, Bulgaria. Association for Computational Linguistics.

Vilar, D., Xu, J., D'Haro, L. F., and Ney, H. (2006). Error analysis of statistical machine translation output. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, pages 697–702.