Towards Automatic Error Analysis of Machine Translation Output

Maja Popović*
RWTH Aachen University

Hermann Ney**
RWTH Aachen University

Evaluation and error analysis of machine translation output are important but difficult tasks. In this article, we propose a framework for automatic error analysis and classification based on the identification of actual erroneous words using the algorithms for computation of Word Error Rate (WER) and Position-independent word Error Rate (PER), which is just a very first step towards development of automatic evaluation measures that provide more specific information of certain translation problems. The proposed approach enables the use of various types of linguistic knowledge in order to classify translation errors in many different ways. This work focuses on one possible set-up, namely, on five error categories: inflectional errors, errors due to wrong word order, missing words, extra words, and incorrect lexical choices. For each of the categories, we analyze the contribution of various POS classes. We compared the results of automatic error analysis with the results of human error analysis in order to investigate two possible applications: estimating the contribution of each error type in a given translation output in order to identify the main sources of errors for a given translation system, and comparing different translation outputs using the introduced error categories in order to obtain more information about advantages and disadvantages of different systems and possibilites for improvements, as well as about advantages and disadvantages of applied methods for improvements. We used Arabic–English Newswire and Broadcast News and Chinese–English Newswire outputs created in the framework of the GALE project, several Spanish and English European Parliament outputs generated during the TC-Star project, and three German–English outputs generated in the framework of the fourth Machine Translation Workshop. We show that our results correlate very well with the results of a human error analysis, and that all our metrics except the extra words reflect well the differences between different versions of the same translation system as well as the differences between different translation systems.

* Now at DFKI – German Research Centre for Artificial Intelligence, Alt-Moabit 91c, 10559 Berlin, Germany. E-mail: maja.popovic@dfki.de.

** Lehrstuhl für Informatik 6 – Computer Science Department, Ahornstrasse 55, 52056 Aachen, Germany. E-mail: ney@informatik.rwth-aachen.de.

Submission received: 8 August 2008; revised submission received: 6 December 2010; accepted for publication: 6 March 2011.

1. Introduction

The evaluation of machine translation output is an important and at the same time difficult task for the progress of the field. Because there is no unique reference translation for a text (as for example in speech recognition), automatic measures are hard to define. Human evaluation, although of course providing (at least in principle) the most reliable judgments, is costly and time consuming. A great deal of effort has been spent on finding measures that correlate well with human judgments when determining which one of a set of translation systems is the best (be it different versions of the same system in the development phase or a set of "competing" systems, as for example in a machine translation evaluation).

However, most of the work has been focused just on best–worst decisions, namely, finding a ranking between different machine translation systems. Although this is useful information and helps in the continuous improvement of machine translation (MT) systems, MT researches often would find it helpful to have additional information about their systems. What are the strengths of their systems? Where do they make errors? Does a particular modification improve some aspect of the system, although perhaps it does not improve the overall score in terms of one of the standard measures? Does a worse-ranked system outperform a best-ranked one in any aspect? Hardly any systematic work has been done in this direction and developers must resort to looking at the translation outputs in order to obtain an insight of the actual problems of their systems. A framework for human error analysis and error classification has been proposed by Vilar et al. (2006), but as every human evaluation, this is also a difficult and time-consuming task.

This article presents a framework for automatic analysis and classification of errors in a machine translation output which is just a very first step in this direction. The basic idea is to extend the standard error rates using linguistic knowledge. The first step is the identification of the actual erroneous words using the algorithms for the calculation of Word Error Rate (WER) and Position-independent word Error Rate (PER). The extracted erroneous words can then be used in combination with different types of linguistic knowledge, such as base forms, Part-of-Speech (POS) tags, Name Entity (NE) tags, compound words, suffixes, prefixes, and so on, in order to obtain various details about the nature of actual errors, for example, error categories (e.g., morphological errors, reordering errors, missing words), contribution of different word classes (e.g., POS, NE), and so forth.

The focus of this work is the definition of the following error categories:

- inflectional errors
- reordering errors
- missing words
- extra words
- incorrect lexical choices

and the comparison of the results of automatic error analysis with those obtained by human error analysis for these categories. Each error category can be further classified according to POS tags (e.g., inflectional errors of verbs, missing pronouns). The translation outputs used for the comparison of human and automatic error analysis

were produced in the frameworks of the GALE[1] project, the TC-STAR[2] project, and the fourth Workshop on Statistical Machine Translation[3] (WMT09). The comparison with human error analysis is done considering two possible applications: estimating the contribution of each error category in a particular translation output, and comparing different translation outputs using these categories. In addition, we show how the new error measures can be used to get more information about the differences between translation systems trained on different source and target languages, between different training set-ups for a same phrase-based translation system, as well as between different translation systems.

1.1 Related Work

A number of automatic evaluation measures for machine translation output have been investigated in recent years. The BLEU metric (Papineni et al. 2002) and the closely related NIST metric (Doddington 2002), along with WER and PER, have been widely used by many machine translation researchers. The Translation Edit Rate (TER) (Snover et al. 2006) and the CDER measure (Leusch, Ueffing, and Ney 2006) are based on the edit distance (WER) but allow reordering of blocks. TER uses an edit distance with additional costs for shifts of word sequences. The CDER measure drops certain constraints for the hypothesis: Only the words in the reference have to be covered exactly once, whereas those in the hypothesis can be covered zero, one, or multiple times. Preprocessing and normalization methods for improving the evaluation using the standard measures WER, PER, BLEU, and NIST are investigated by Leusch et al. (2005). The same set of measures is examined by Matusov et al. (2005) in combination with automatic sentence segmentation in order to enable evaluation of translation output without sentence boundaries (e.g., translation of speech recognition output). The METEOR metric (Banerjee and Lavie 2005) first counts the number of exact word matches between the output and the reference. In a second step, unmatched words are converted into stems or synonyms and then matched. A method that uses the concept of maximum matching string (MMS) is presented by Turian, Shen, and Melamed (2003). IQ (Giménez and Amigó 2006) is a framework for automatic evaluation in which evaluation metrics can be combined. Nevertheless, none of these measures or extensions takes into account any details about actual translation errors, for example, what the contribution of verbs is in the overall error rate, how many full forms are wrong although their base forms are correct, or how many words are missing. A framework for human error analysis and error classification has been proposed by Vilar et al. (2006), where a classification scheme (Llitjós, Carbonell, and Lavie 2005) is presented together with a detailed analysis of the obtained results.

Automatic error analysis is still a rather unexplored area. A method for automatic identification of patterns in translation output using POS sequences is proposed by Lopez and Resnik (2005) in order to see how well a translation system is capable of capturing systematic reordering patterns. Using relative differences between WER and PER for three POS classes (nouns, adjectives, and verbs) is proposed by Popović et al. (2006) for the estimation of inflectional and reordering errors. Semi-automatic error analysis (Kirchhoff et al. 2007) is carried out in order to identify problematic

1 GALE — Global Autonomous Language Exploitation. http://www.arpa.mil/ipto/programs/gale/index.htm.
2 TC-STAR — Technology and Corpora for Speech to Speech Translation. http://www.tc-star.org/.
3 EACL 09 Fourth Workshop on Statistical Machine Translation. http://www.statmt.org/wmt09/.

characteristics of source documents such as genre, domain, language, and so on. Zhou et al. (2008) propose a diagnostic evaluation of linguistic check-points obtained automatically by aligning parsed source and target sentences. For each check-point, the number of matched *n*-grams of the references is then calculated. Linguistically based reordering along with the syntax-based evaluation of reordering patterns is described in Xiong et al. (2010).

In this work, we propose a novel framework for automatic error analysis of machine translation output based on WER and PER, and systematically investigate a set of possible methods to carry out an error analysis at the word level.

2. A Framework for Automatic Error Analysis

The basic idea for automatic error analysis described in this work is to take into account details from the WER (edit distance) and PER algorithms, namely, to identify all erroneous words which are actually contributing to the error rate, and then to combine these words with different types of linguistic knowledge. The general procedure for automatic error analysis and classification is shown in Figure 1. An overview of the standard error rates WER and PER is given in Section 2.1, and methods for extracting actual errors are described in the following sections.

In this article, we carried out the error analysis at the word level and we used base forms of the words and POS tags as linguistic knowledge. However, the analysis described in this work is just a first step towards automatic error analysis and presents only one of many possibilities—this framework enables the integration of various knowledge sources such as deeper lingustic knowledge, the introduction of source words (possibly with additional linguistic information) if appropriate alignment information is available, and so forth. Investigation at the word group/phrase level instead of only at the word level is possible as well. The error analysis presented in this

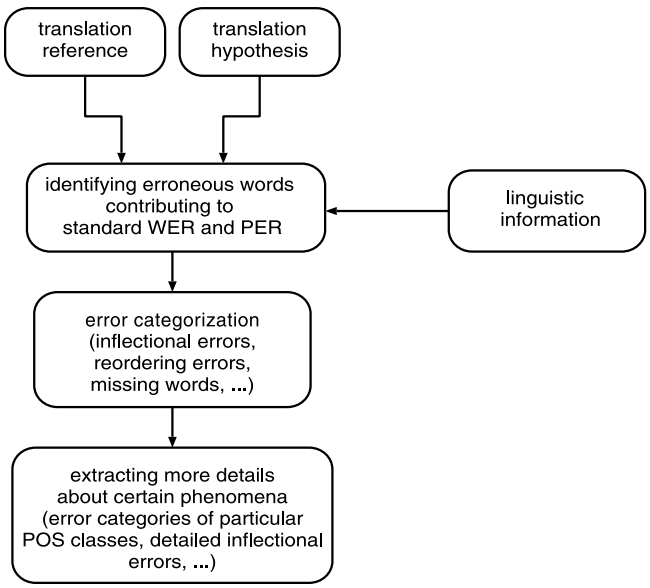

Figure 1
General procedure for automatic error analysis based on the standard word error rates and linguistic information.

work is language-independent—nevertheless, availability of base forms and POS tags for the particular target language is a requisite.

2.1 Standard Word Error Rates (Overview)

The standard procedure for evaluating machine translation output is done by comparing the hypothesis document *hyp* with the given reference document *ref*, each one consisting of K sentences (or segments). The reference document *ref* consists of $N_R \geq 1$ reference translations of the source text. $N_R = 1$ stands for the case when only a single reference translation is available, and $N_R > 1$ denotes the case of multiple references. Let the length of the hypothesis sentence hyp_k be denoted as N_{hyp_k}, and the length of each reference sentence $N_{ref_{k,r}}$. Then, the total hypothesis length of the document is $N_{hyp} = \sum_k N_{hyp_k}$ and the total reference length is $N_{ref} = \sum_k N^*_{ref_k}$, where $N^*_{ref_k}$ is defined as the length of the reference sentence with the lowest sentence-level error rate as shown to be optimal with respect to the correlation with the human evaluation score's adequacy and fluency (Leusch et al. 2005). The overall error rate is then obtained by normalizing the total number of errors over the total reference length.

The word error rate (WER) is based on the Levenshtein distance (Levenshtein 1966)—the minimum number of substitutions, deletions, and insertions that have to be performed to convert the generated text *hyp* into the reference text *ref*. A shortcoming of the WER is the fact that it does not allow reorderings of words, although the word order of the hypothesis can be different from the word order of the reference even though it is a correct translation. The position-independent word error rate (PER) is also based on substitutions, deletions, and insertions but without taking the word order into account. The PER is always lower than or equal to the WER. On the other hand, a shortcoming of the PER is the fact that it does not penalize a wrong word order.

Calculation of WER: The WER of the hypothesis *hyp* with respect to the reference *ref* is calculated as

$$\text{WER} = \frac{1}{N_{ref}} \sum_{k=1}^{K} \min_r \left\{ d_L(ref_{k,r}, hyp_k) \right\} \tag{1}$$

where $d_L(ref_{k,r}, hyp_k)$ is the Levenshtein distance between the reference sentence $ref_{k,r}$ and the hypothesis sentence hyp_k. The calculation is performed using a dynamic programming algorithm.

Calculation of PER: Define $n(w, setw)$ as the number of occurrences of a word w in a multi-set of words $setw$. The PER can be calculated using the counts $n(e, hyp_k)$ and $n(e, ref_{k,r})$ of a word e in the hypothesis sentence hyp_k and the reference sentence $ref_{k,r}$ respectively:

$$\text{PER} = \frac{1}{N_{ref}} \sum_{k=1}^{K} \min_r \left\{ d_{\text{PER}}(ref_{k,r}, hyp_k) \right\} \tag{2}$$

where

$$d_{\text{PER}}(ref_{k,r}, hyp_k) = \frac{1}{2} \left(|N_{ref_{k,r}} - N_{hyp_k}| + \sum_e |n(e, ref_{k,r}) - n(e, hyp_k)| \right) \tag{3}$$

2.2 Identification of WER Errors

The dynamic programming algorithm for WER enables a simple and straightforward identification of each erroneous word which actually contributes to WER. An example of a reference sentence and hypothesis sentence along with the corresponding Levenshtein alignment and the actual words participating in WER is shown in Table 1. The reference words involved in WER are denoted as reference errors, and hypothesis errors refer to the hypothesis words participating in WER.

Table 2 presents an example of introducing linguistic knowledge in the form of base forms and POS tags. This allows us to compute the contribution of each POS class p to the overall WER, that is, WER(p). If $werr_k$ is the multi-set of erroneous words in sentence k according to alignment links to the best reference and p is a POS class, then $N(\text{WER}(\text{p})) = \sum_{e \in \text{p}} n(e, werr_k)$ is the number of WER errors in $werr_k$ produced by words belonging to the POS class p. For the substitution and the deletion errors, POS tags of

Table 1
Example for illustration of actual errors: (a) a reference sentence and a corresponding hypothesis sentence; (b) a corresponding Levenshtein alignment; (c) actual words which participate in the word error rate.

(a) Reference and hypothesis example

reference:	hypothesis:
Mister Commissioner, twenty-four hours sometimes can be too much time.	Mrs Commissioner, sometimes twenty-four hours is too much time.

(b) Levenshtein alignment

ref:	hyp:
Mister	Mrs
Commissioner	Commissioner
,	,
	sometimes
twenty-four	twenty-four
hours	hours
sometimes	is
can	
be	
too	too
much	much
time	time
.	.

(c) WER errors

reference errors	hypothesis errors	error type
Mister	Mrs	substitution
	sometimes	insertion
sometimes	is	substitution
can		deletion
be		deletion

Table 2
WER errors and linguistic knowledge: Actual words which participate in the word error rate with their corresponding base forms and POS classes.

reference errors	hypothesis errors	error type
Mister#Mister#N	Mrs#Mrs#N	substitution
	sometimes#sometimes#ADV	insertion
sometimes#sometimes#ADV	is#be#V	substitution
can#can#V		deletion
be#be#V		deletion

the reference words are used, and for the insertion errors, POS classes of the hypothesis words are used. The WER for the word class p can be calculated as the standard WER by normalizing the number of errors over the total reference length:

$$\text{WER}(\text{p}) = \frac{1}{N_{ref}} \sum_{k=1}^{K} \sum_{e \in \text{p}} n(e, werr_k) \tag{4}$$

Standard WER of the whole sentence is equal to $5/12 = 41.7\%$. The contribution of nouns is $\text{WER}(\text{N}) = 1/12 = 8.3\%$, of verbs is $\text{WER}(\text{V}) = 2/12 = 16.7\%$, and of adverbs is $\text{WER}(\text{ADV}) = 2/12 = 16.7\%$.

2.3 Identification of PER Errors

In contrast to WER, the standard efficient algorithms for the calculation of PER do not give precise information about contributing words. However, it is possible to identify all words in the hypothesis which do not have a counterpart in the reference, and vice versa. These words will be referred to as PER errors.

An illustration of PER errors is given in Table 3. The number of errors contributing to the standard PER according to Equation (3) is 3—there are two substitutions and one deletion. The problem with standard PER is that it is not possible to detect which words are deletion errors, which are insertion errors, and which words are substitution errors. Therefore we introduce alternative PER-based measures which correspond to the precision, recall, and F–measure. Let $herr_k$ refer to the multi-set of words in the hypothesis sentence k which do not appear in the reference sentence k (referred to as **hypothesis errors**). Analogously, let $rerr_k$ denote the multi-set of words in the reference

Table 3
PER errors: Actual words that participate in the position-independent word error rate.

reference errors	hypothesis errors
Mister	Mrs
can	is
be	

sentence k which do not appear in the hypothesis sentence k (referred to as **reference errors**). Then the following measures can be calculated:

- recall-based (reference) PER (RPER):

$$\text{RPER} = \frac{1}{N_{ref}} \sum_{k=1}^{K} \sum_{e} n(e, rerr_k) \tag{5}$$

- precision-based (hypothesis) PER (HPER):

$$\text{HPER} = \frac{1}{N_{hyp}} \sum_{k=1}^{K} \sum_{e} n(e, herr_k) \tag{6}$$

- F-based PER (FPER):

$$\text{FPER} = \frac{1}{N_{ref} + N_{hyp}} \cdot \sum_{k=1}^{K} \sum_{e} \Big(n(e, rerr_k) + n(e, herr_k) \Big) \tag{7}$$

For the example sentence presented in Table 1, the number of hypothesis errors $\sum_e n(e, herr_k)$ is 2 and the number of reference errors $\sum_e n(e, rerr_k)$ is 3. The number of errors contributing to the standard PER is 3 according to Equation (3), since $|N_{ref} - N_{hyp}| = 1$ and $\sum_e |n(e, ref_k) - n(e, hyp_k)| = 5$. The standard PER is normalized over the reference length $N_{ref} = 12$, thus being equal to 25%. The RPER considers only the reference errors, $\text{RPER} = 3/12 = 25\%$, and HPER only the hypothesis errors, $\text{HPER} = 2/11 = 18.2\%$. The FPER is the sum of hypothesis and reference errors divided by the sum of hypothesis and reference length: $\text{FPER} = (2 + 3)/(11 + 12) = 5/23 = 21.7\%$.

The contribution of nouns in the reference translation is $\text{RPER(N)} = 1/12 = 8.3\%$, in the hypothesis is $\text{HPER(N)} = 1/11 = 9.1\%$, and together $\text{FPER(N)} = 2/23 = 8.7\%$. The contribution of verbs in the reference is $\text{RPER(V)} = 2/12 = 16.7\%$, in the hypothesis is $\text{HPER(V)} = 1/11 = 9.1\%$, and together $\text{FPER(V)} = 3/23 = 13\%$.

It should be noted that only the links between raw words are considered both for WER as well as for RPER and HPER calculation; POS tags are added afterwards as an additional knowledge. Exact distribution of errors over POS classes depends on the exact implementation of the WER and the RPER and HPER algorithms. For example, if *light#A* and *light#N* occur in the reference and *light#V* occurs in the hypothesis, there are two possibilities: either *light#A* is linked to *light#V* and *light#N* is a missing word, or *light#N* is linked to *light#V* so that *light#A* is a missing word.

3. Methods for Automatic Error Analysis and Classification

The error details described in Section 2.2 and Section 2.3 can be combined with different types of linguistic knowledge in different ways. Examples with the base forms and POS tags as linguistic knowledge are presented in Tables 2 and 4. The described error rates

Table 4
PER errors and linguistic knowledge: Actual words that participate in the position-independent word error rate and their corresponding base forms and POS classes.

reference errors	hypothesis errors
Mister#Mister#N	Mrs#Mrs#N
can#can#V	is#be#V
be#be#V	

of particular POS classes give more details than the overall standard error rates and can be used for error analysis to some extent. However, for more precise information about certain phenomena some kind of further analysis is required. In this work, we examine the following error categories:

- inflectional errors — using RPER or HPER errors and base forms;
- reordering errors — using WER and RPER or HPER errors;
- missing words — using WER and RPER errors with base forms;
- extra words — using WER and HPER errors with base forms;
- incorrect lexical choice — errors which belong neither to inflectional errors nor to missing or extra words.

Furthermore, the contribution of various POS classes for the described error categories is estimated.

It should be noted that the base forms and POS tags are needed both for the reference(s) and for the hypothesis. The performance of morpho-syntactic analysis is slightly lower on the hypothesis, but this does not seem to influence the performance of the error analysis tools. We choose to use reference words for all cases where it can be chosen between the reference and the hypothesis, however. Nevertheless, it would be interesting to investigate the use of hypothesis words in future experiments and compare the results.

3.1 Inflectional Errors

An inflectional error occurs if the base form of the generated word is correct but the full form is wrong. Inflectional errors can be estimated using RPER errors and base forms in the following way: From each reference–hypothesis sentence pair, only erroneous words which have common base forms are taken into account:

$$N(infl) = \sum_{k=1}^{K} \sum_{e} n(e, rerr_k) - \sum_{k=1}^{K} \sum_{eb} n(eb, rberr_k) \tag{8}$$

where *eb* denotes the base form of the word *e* and $rberr_k$ stands for the multi-set of base form errors in the reference. The number of words with erroneous base forms (representing a multi-set of non–inflectional errors) is subtracted from the number of total errors. For example, from the PER errors presented in Table 3, the word *is* will be

detected as an inflectional error because it shares the same base form with the reference error *be*.

An analogous definition is possible using HPER errors and base forms; as explained at the beginning of this section, however, we choose to use the reference words because the results of the morpho-syntactic analysis are slightly more reliable for the references than for the hypotheses.

3.2 Reordering Errors

Differences of word order in the hypothesis with respect to the reference are taken into account only by WER and not by PER. Therefore, a word which occurs both in the reference and in the hypothesis but is marked as a WER error is considered as a reordering error. The contribution of reference reordering errors can be estimated in the following way:

$$N(reord) = \sum_{k=1}^{K} \sum_{e} \Big(n(e, suberr_k) + n(e, delerr_k) - n(e, rerr_k) \Big) \tag{9}$$

where $suberr_k$ represents the multi-set of WER substitution errors, $delerr_k$ the multi-set of WER deletion errors, and $rerr_k$ the multi-set of RPER errors. A definition using HPER errors with substitutions and insertions is also possible; this work, however, is focused on the reference errors. For the example in Table 1, the word *sometimes* is identified as a reordering error.

3.3 Missing Words

Missing words can be identified using the WER and PER errors in the following way: The words considered as missing are those which occur as deletions in WER errors and at the same time occur only as reference PER errors without sharing the base form with any hypothesis error, that is, as a non–inflectional RPER error:

$$N(miss) = \sum_{k=1}^{K} \sum_{eb \in rberr_k} n(e, delerr_k) \tag{10}$$

The multi-set of deletion WER errors is defined as $delerr_k$, and $rberr_k$ stands for the multi-set of base form RPER errors. The use of both WER and RPER errors is much more reliable than using only the WER deletion errors because not all deletion errors are produced by missing words—a number of WER deletions appear due to reordering errors. The information about the base form is used in order to eliminate inflectional errors. For the example in Table 1, the word *can* will be identified as missing.

3.4 Extra Words

Analogously to missing words, extra words are also detected from the WER and PER errors: The words considered as extra are those that occur as insertions in WER errors

and at the same time occur only as hypothesis PER errors without sharing the base form with any reference error:

$$N(extra) = \sum_{k=1}^{K} \sum_{eb \in hberr_k} n(e, inserr_k) \tag{11}$$

where $inserr_k$ is the multi-set of insertion WER errors and $hberr_k$ is the multi-set of base form HPER errors. In the example in Table 1 none of the words will be classified as an extra word.

3.5 Incorrect Lexical Choice

The erroneous words in the reference translation that are classified neither as inflectional errors nor as missing words are considered as incorrect lexical choice:

$$N(lex) = \sum_{k=1}^{K} \sum_{eb} n(eb, rberr_k) - N(miss) \tag{12}$$

As in the case of the inflectional and reordering errors, a definition using hypothesis errors and extra words is also possible, but in this work we choose to use reference errors. In the example in Table 1, the word *Mister* in the reference (or the word *Mrs* in the hypothesis) is considered as an incorrect lexical choice.

4. Comparison with Human Error Analysis

In order to compare the results of the proposed automatic error analysis with human error analysis, the methods described in the previous sections are applied on several translation outputs with the available results of human error analysis. These translation outputs were produced in the framework of the GALE project, the TC-STAR project, and the shared task of the fourth Statistical Machine Translation Workshop (WMT09).

The two main goals of the comparison with the human evaluation are:

- to examine the distribution of errors over the categories, that is, how well the automatic methods are capable of capturing differences between error categories and determining which of those are particularly problematic for a given translation system;
- to examine the differences between the numbers of errors in each category for different translation outputs, that is, how well the automatic methods are capable of capturing differences between systems.

4.1 Human Error Analysis

Human error analysis and classification is a time-consuming and difficult task, and it can be done in various ways. For example, in order to find errors in a translation output it can be useful to have one or more reference translations. There are often several correct translations of a given source sentence, however, and some of them might not correspond to the reference translations, which poses difficulties for evaluation and

error analysis. The errors can be counted by doing a direct strict comparison between the references and the translation outputs, which is then very similar to automatic error analysis. But much more flexibility can be allowed: substitution of words and expressions by synonyms, syntactically correct different word order, and so on, which is a more natural way. It is also possible to use the references only for the semantic aspect, namely, to look only whether the main meaning is preserved. It is even possible not to use a reference translation at all, but compare the translation output with the source text. There are also other aspects that may differ between human evaluations, for example, counting each problematic word as an error or counting groups of words as one error, and so forth. Furthermore, the human error classification is definitely not unambigous—often it is not easy to determine in which particular error category some error exactly belongs, sometimes one word can be assigned to more than one category, and variations between different human evaluators are possible. For error categories described in previous sections, especially difficult is disambiguating between incorrect lexical choice and missing words or extra words. For example, if the translation output is *the day before yesterday* and translation reference is *yesterday*, it could be considered as a group of incorrectly translated words, but also as a group of extra words. Similarly, there are several possible interpretations of errors if *the one who will come* is translated as *which comes*.

In this work, three types of human error analysis are used:

- a strict one, comparing the output with a given reference (similar to the automatic error analysis) (Table 5);
- a flexible one, where syntactically correct differences in word order, substitutions by synonyms, and correct alternative expressions are not considered as errors; less strict than the previous method (Table 5);
- a free one, where the reference is taken into account only from the semantic point of view (Vilar et al. 2006); less strict than the previous two methods.

The results of both human and automatic error analysis for all analyzed texts are presented in the following sections. In addition, the Pearson (r) and Spearman rank (ρ) correlation coefficients between human and automatic results are calculated. Both coefficients assess how well a monotonic function describes the relationship between two variables: The Pearson correlation assumes a linear relationship between the variables, and the Spearman correlation takes only rank into account. Thus Spearman's

Table 5
Examples of two variants of human error analysis, a strict and a flexible one; the marked errors are detected with respect to the reference, whereas no errors are detected when the error analysis is more flexible.

reference translation	obtained output
we celebrated the fifteenth anniversary	we **have held** the fifteenth anniversary
I think this is a good moment	I **believe that** this is a good **opportunity**
to achieve these ends	**for** these **purposes**
in 2002	in **the year** 2002
in Europe we must also learn	**also** in Europe we must learn

rank correlation coefficient is equivalent to a Pearson correlation on ranks. A Pearson correlation of +1 means that there is a perfect positive linear relationship between the variables, and a Spearman correlation of +1 that the ranking using both variables is exactly the same. A Pearson correlation of −1 means that there is a perfect negative linear relationship between variables, and a Spearman correlation of −1 that there is an exactly inverse ranking. A correlation of 0 means there is no linear relationship between the two variables. Thus, the higher value of r and ρ, the more similar the metrics are.

4.2 Distribution of Errors Over Categories

The goal of the experiments described in this section is to examine the distribution of errors over the categories, that is, how well the automatic methods are capable of capturing differences between error categories and determining which of those are particularly problematic for a given translation system. For each of the error categories, the distribution of errors over the basic POS classes—nouns (N), verbs (V), adjectives (A), adverbs (ADV), pronouns (PRON), determiners (DET), prepositions (PREP), conjunctions (CON), numerals (NUM), and punctuation marks (PUN)—is analyzed as well, thus obtaining more details about errors, namely, which POS classes are the main source of errors for a particular error category. For the GALE corpora, the strict human error analysis is carried out, and for the TC-STAR corpora, the free one.

4.2.1 Results on GALE Corpora. The raw error counts for each category obtained on the GALE corpora both by human and automatic error classification are shown in Table 6. It can be seen that both the results of the human analysis as well as the automatic analysis show the same tendencies: For the Arabic-to-English Broadcast News translation, the main sources of errors are extra words and incorrect lexical choice, for the Newswire corpus the predominant problem is incorrect lexical choice, and for the Chinese-to-English the majority of errors are caused by missing words, followed by incorrect lexical choices and wrong word order.

Three examples of human and automatic error analysis are presented in Table 7. In the first sentence, the words *Japanese* and *friendly* are classified into the same category both by human and by automatic analysis, namely, as a reordering error and a missing word, respectively. The words *feeling for* represent an example where the human analysis assigns the error to the category of missing words, but the automatic analysis classifies it as a lexical error. Similarly, the words *can feel* are considered as extra words by humans, but as lexical errors by automatic tools. These examples illustrate the previous statements about difficulties in disambiguation between missing words and extra words vs. lexical errors. In the second sentence, the inflectional error *based/base* is detected both by humans and by automatic tools. However, *contribution* is classified

Table 6
Results (raw error counts) of human (left) and automatic (right) error analysis for the GALE corpora.

output	infl	order	miss	ext	lex
ArEn BN	20 / 23	39 / 66	79 / 63	127 / 137	135 / 147
ArEn NW	22 / 24	30 / 41	97 / 102	73 / 76	140 / 131
CnEn NW	38 / 40	127 / 171	288 / 244	95 / 117	203 / 239

Table 7
Examples of human and automatic error analysis from the GALE corpora. Words in ***bold italic*** are assigned to the same error category both by human and automatic error analysis, and words in **bold** represent differences.

ref:	... , although the ***Japanese friendly*** **feelings for** China added an increase , ...
hyp:	... , although China **can feel** the Japanese increase , ...
errors:	friendly – missing(hum,aut) feelings for – missing(hum)/lexical(aut) Japanese – order(hum,aut) can feel – extra(hum)/lexical(aut)
ref:	...the amount of their monthly **contribution** is ***based*** in accordance with the **wages** of previous year...
hyp:	...the amount of their monthly wages base in accordance with the previous year...
errors:	contribution – lexical(hum)/missing(aut) wages – missing(hum)/order(aut) based – inflectional(hum,aut)
ref:	... of local party committees. ***Secretaries*** of the Commission ...
hyp:	... of local party committees of the ***provincial*** Commission ...
errors:	Secretaries – missing(hum,aut) provincial – extra(hum,aut)

as a lexical error by humans and as a missing word by automatic tools. The human analysis classified the word *wages* as missing. The word is detected as an error also by the automatic tool; nevertheless it is considered as a reordering error because it is present only as an WER error and neither as an HPER nor as an RPER error. The third sentence illustrates a total agreement between the human and automatic error classification: Both words are assigned to the same category. Results for the ten basic POS classes are shown in Table 8, and again from both human and automatic error analysis the same conclusions can be drawn.

Table 9 presents correlations between the results of the human and automatic analysis. The correlation function presented in Table 9(a) is measured between the error counts in each category, and Table 9(b) presents the correlation between the error counts for each POS class within a particular error category. It can be seen that the automatic measures have very high correlation coefficients with respect to the results of human evaluation. The correlations for the inflectional error category are higher than for the other categories, which can be explained by the fact mentioned in previous sections that the disambiguation between missing words, extra words, and incorrect lexical choice is often difficult, both for humans and for machines.

4.2.2 Results on TC-STAR Corpora. The experiments on the TC-STAR corpora are similar to those on the GALE corpora. There are some differences, however, because human error classification is carried out in a somewhat different way and completely independently. The error categories considered by the human error analysis were inflectional errors, missing words, reordering errors, and incorrect lexical choice—that is, the same as in the GALE experiments except extra words. The distribution of errors over POS tags is not analyzed on this corpora, but the following details about inflectional errors are investigated: verb tense errors, verb person errors, adjective gender errors, and adjective

Table 8
Results (raw error counts) of human (left) and automatic (right) error analysis for the GALE corpora: Distribution of different error types over basic POS classes.

ArEn BN	V	N	A	ADV	PRON	DET	PREP	CON	NUM	PUN
infl	15 / 17	3	0	0	2 / 3	0	0	0	0	0
order	6 / 10	14 / 15	3 / 6	1 / 2	0	5 / 10	3 / 8	3 / 7	2 / 1	1 / 6
miss	29 / 15	10 / 14	4 / 2	6 / 4	11 / 9	3 / 2	8 / 5	1	1 / 0	6 / 11
ext	11	19 / 23	4 / 3	9 / 8	8 / 15	33 / 36	25 / 21	7 / 4	3 / 5	8 / 11
lex	22 / 32	23 / 22	7	7 / 9	17 / 16	5	24	10 / 7	6	14 / 19

ArEn NW	V	N	A	ADV	PRON	DET	PREP	CON	NUM	PUN
infl	18	2	1	0	1 / 3	0	0	0	0	0
order	5	9 / 10	4 / 3	5	0 / 1	2 / 3	3 / 9	1 / 2	1	0 / 2
miss	25 / 35	14 / 12	4 / 3	5 / 4	18 / 14	9 / 10	15 / 14	5 / 3	0	2 / 7
ext	8 / 11	12 / 17	2	1 / 3	7 / 4	10 / 8	12 / 10	5	0	15 / 16
lex	38 / 27	24 / 22	4 / 7	8 / 9	22 / 23	8 / 7	23 / 17	7 / 10	2	4 / 7

CnEn NW	V	N	A	ADV	PRON	DET	PREP	CON	NUM	PUN V
infl	14 / 16	24	0	0	0	0	0	0	0	0
order	12 / 13	52 / 71	16 / 12	3 / 2	4 / 1	12 / 22	16 / 22	3 / 5	4 / 5	5 / 18
miss	49 / 45	75 / 73	12 / 7	14 / 13	17 / 13	24 / 15	50 / 36	23 / 13	5	19 / 22
ext	6	18 / 38	5 / 9	1 / 0	2 / 1	21 / 20	23 / 24	5 / 2	0 / 4	14 / 13
lex	21 / 47	87 / 72	13	7 / 11	5 / 11	13 / 10	32 / 37	9 / 15	4 / 6	12 / 19

number errors. The category of inflectional errors is also different: It is obtained as a sum of these particular inflectional categories. Correlation coefficients are calculated both for general error categories and for inflectional details.

The results of this error classification are shown in Table 10, and it can be seen that human and automatic error analysis again produce similar trends. It can be seen as

Table 9
Correlation coefficients for the GALE corpora: Spearman rank ρ (left column) and Pearson *r* (right column) coefficient.

(a) Error categories

output	ρ	*r*
ArEn BN	0.900	0.955
ArEn NW	1.000	0.994
CnEn NW	1.000	0.930

(b) Distribution of errors over POS classes

	distribution of errors over POS classes									
	infl		order		miss		extra		lex	
output	ρ	*r*	ρ	*r*	ρ	*r*	ρ	*r*	ρ	*r*
ArEn BN	0.997	0.999	0.927	0.872	0.918	0.790	0.870	0.947	0.924	0.924
ArEn NW	0.979	0.994	0.870	0.804	0.921	0.922	0.912	0.916	0.894	0.960
CnEn NW	1.000	0.998	0.812	0.961	0.927	0.973	0.879	0.853	0.788	0.914

Table 10
Results (raw error counts) of human (left) and automatic (right) error analysis for the TC-STAR corpora.

output	infl	order	miss	lex
EsEn1 FTE	18 / 77	37 / 156	47 / 138	70 / 336
EsEn2 FTE	27 / 136	28 / 215	46 / 154	47 / 477
EsEn1 VT	24 / 80	43 / 104	40 / 113	82 / 268
EnEs1 FTE	89 / 264	45 / 194	58 / 206	89 / 451
EnEs2 FTE	72 / 197	33 / 169	38 / 111	64 / 416
EnEs1 VT	82 / 223	31 / 178	72 / 187	84 / 485

well that the numbers of errors obtained by automatic methods is much higher than the numbers obtained by the free human evaluation.

Table 11 presents the results for the inflectional details about verbs and adjectives (i.e., tense, person, gender, and number). Both human and automatic error analysis indicate that the most problematic inflectional category is the tense of verbs, especially for the translation into Spanish.

Correlation coefficients are shown in Table 12. It can be seen that for this corpus the correlations for the error categories, although all rather high (above 0.5), are lower than for the GALE corpus. This is due to the free human evaluation which is carried out on this corpora, that is, without taking the reference translation strictly into account. However, for the inflectional error analysis the correlations are very high, above 0.9.

4.3 Differences Between Translation Systems and Methods for Improvements

The focus of this set of experiments is to examine how well the automatic methods are capable of capturing differences between systems and methods for improvements in order to

- estimate advantages and disadvantages of different translation systems;
- get ideas for improvements of translation performance;
- estimate advantages and disadvantages of applied methods for improvements.

Table 11
Results (raw error counts) of human (left) and automatic (right) error analysis for TC-STAR corpora — inflectional details: tense (Vten) and person (Vper) of verbs, gender (Agen), and number (Anum) of adjectives.

output	Vten	Vper	Agen	Anum
EsEn1 FTE	14 / 56	4 / 21	0	0
EsEn2 FTE	22 / 98	5 / 38	0	0
EsEn1 VT	16 / 59	8 / 21	0	0
EnEs1 FTE	44 / 131	24 / 82	12 / 26	9 / 25
EnEs2 FTE	31 / 94	18 / 52	12 / 26	11 / 25
EnEs1 VT	36 / 120	23 / 75	13 / 14	10 / 14

Table 12
Correlation coefficients for the TC-STAR corpora: Spearman rank ρ (left column) and Pearson *r* (right column).

output	error categories		infl. errors over POS classes	
	ρ	*r*	ρ	*r*
EsEn1 FTE	0.800	0.935	1.000	0.996
EsEn2 FTE	0.800	0.552	1.000	0.983
EsEn1 VT	0.800	0.978	1.000	0.991
EnEs1 FTE	0.950	0.754	1.000	0.987
EnEs2 FTE	0.600	0.572	1.000	0.998
EnEs1 VT	1.000	0.538	0.950	0.990

The experiments should also show how reliable each error category is for the comparison of translation outputs. For each of the error categories, the basic POS classes are analyzed as well in order to estimate which POS classes of each category are reliable for the comparison of translation outputs.

The investigations are carried out on six Spanish-to-English TC-STAR outputs generated by phrase-based systems (Vilar et al. 2005) and three German-to-English WMT09 outputs produced in the framework of the fourth shared translation task (Callison-Burch et al. 2009). Two of the WMT09 outputs are generated by standard phrase-based systems (Zens, Och, and Ney 2002) and one by a hierarchical phrase-based system (Chiang 2007). For the TC-STAR outputs two reference translations are available for the automatic error analysis, and for the WMT09 outputs only a single reference is available. For all texts, the flexible human error analysis is carried out. The following sections summarize all the results along with the Spearman and Pearson correlation coefficients calculated across the different translation outputs.

4.3.1 Results on TC-STAR Corpora. The error analyses were carried out on six Spanish-to-English outputs generated by phrase-based translation systems built on different sizes of training corpora in order to examine the effects of data sparseness (Popović and Ney 2006b). In addition, the effects of local POS-based word reorderings of nouns and adjectives (Popović and Ney 2006a) were analyzed in order to examine improvements of the baseline system. Adjectives in the Spanish language are usually placed after the corresponding noun, whereas for English it is the other way around. Therefore, local reorderings of nouns and adjective groups in the source language were applied. If the source language is Spanish, each noun is moved behind the corresponding adjective group. If the source language is English, each adjective group is moved behind the corresponding noun. An adverb followed by an adjective (e.g., *more important*) or two adjectives with a coordinate conjunction in between (e.g., *economic and political*) are treated as an adjective group. Reorderings were applied in the source language, then training and search were performed using the transformed source language data. Modifications of the training and search procedure were not necessary. In this work, only Spanish-to-English translation is analyzed. The English outputs of following training set-ups for the same phrase-based translation system are examined:

- training on the full bilingual corpus — a large task–specific corpus (1.3M)

- training on a small task–specific corpus (13k)
- training only on a conventional dictionary

The effects of local reorderings are investigated for each size of the training corpus.

Table 13 presents the raw error counts in each category for each of the six translation outputs in the form N_{hum}/N_{aut}. In the last row of the table, the Spearman and Pearson correlation coefficients ρ and r for each category across different translation outputs are shown. In addition, the correlations for error distributions within a translation output (as in Section 4.2) are presented as well in the rightmost column, and it can be seen that, as in the previous experiments, they are high for each of the translation outputs. As for the correlations of error categories across translation outputs, the class of inflectional errors and of incorrect lexical choice have very high correlations, which suggests that these two categories reflect well the differences between translation outputs. The reordering errors and missing words are also suitable for comparison, whereas the category of extra words has even a negative correlation—it is not possible to draw conclusions about differences between translation outputs looking into this category.

From both the human and the automatic error analysis it can be concluded that the number of inflectional errors is low, and becomes slightly higher for the system trained on a dictionary, although there is a number of reordering errors, missing words, and extra words. The most problematic category is the incorrect lexical choice, especially for the small training corpora. The number of reordering errors and missing words is also increasing when the corpus size is decreasing, although to a lesser extent. The POS-based local reordering technique reduces the number of reordering errors, especially for the small training corpora, and does not harm the other error categories.

4.3.2 Results on WMT Corpora. Three outputs generated by three German-to-English statistical translation systems are analyzed in order to examine the differences between a standard phrase-based translation model and a hierarchical translation model. In the standard phrase-based model, phrases are defined as non-empty contiguous sequences of words. The hierarchical phrase-based model is an extension of this model where the phrases are allowed to have "gaps" (i.e., non-contiguous parts of the source sentence are allowed to be translated into possibly non-contiguous parts of the target sentence).

Table 13
Results (raw error counts) of human (left) and automatic (right) error analysis for six different Spanish-to-English TC-STAR systems; Spearman (left) and Pearson (right) correlation coefficients for each translation output across error categories (last column) and for each error category across different translation outputs (last row).

system	infl	order	miss	extra	lex	ρ/r
1.3M	7 / 11	24 / 47	38 / 48	31 / 37	64 / 184	0.9 / 0.92
+reord. adj.	7 / 11	16 / 46	37 / 49	25 / 31	61 / 184	0.9 / 0.91
13k	7 / 12	35 / 74	55 / 57	26 / 34	134 / 223	0.9 / 0.98
+reord. adj.	8 / 13	19 / 50	54 / 60	25 / 40	132 / 217	0.9 / 0.98
dictionary	14 / 18	45 / 88	104 / 84	33 / 30	378 / 414	0.9 / 0.99
+reord. adj.	14 / 19	21 / 64	104 / 80	35 / 32	375 / 391	0.9 / 0.99
ρ/r	0.94 / 0.99	0.83 / 0.85	0.87 / 0.99	−0.19 / −0.34	0.99 / 0.99	

In this way, long-distance dependencies and reorderings can be modelled better. In addition, we investigate the effects of long range POS-based reorderings of German verbs (Popović and Ney 2006a) used to improve the phrase-based system. Verbs in the German language can often be placed at the end of a clause. This is mostly the case with infinitives and past participles, but there are many cases when other verb forms also occur at the clause end. For the translation from German into English, the following verb types were moved towards the beginning of a clause: infinitives, infinitives with the German infinitive particle *zu*, finite verbs, past participles, and verb negations. Analogously to the local reorderings described in the previous section, training and search are performed using the transformed source language data.

Table 14 presents the raw error counts in the form N_{hum}/N_{aut} for all categories and all translation outputs. Spearman and Pearson correlation coefficients are shown as well for each translation output across error categories (rightmost column), and for each error category across different translation outputs (last row). Again, the correlations across the error categories are very high for all translation outputs, although slightly smaller than for the TC-STAR data. It can be seen that the extra words also have the weakest correlation among error categories, although the coefficients are not negative as in the case of the TC-STAR data. This confirms the previous hypothesis that this error category is not suitable for looking into details about differences between translation outputs.

From the results of both error analyses it can be seen that for all translation outputs the inflectional errors are again the least problematic category. On the other hand, there is a large number of reordering errors, missing words, and especially lexical errors. The main advantage of the hierarchical system compared to the standard phrase-based system is the smaller number of missing words, and the main disadvantage the larger number of reordering errors. This looks rather surprising at first, because the hierarchical system should actually better deal with various reorderings, both local and long-range. Nevertheless, the produced reorderings are rather unconstrained so that a number of them are beneficial, but there are also a number of reorderings that make the translation quality worse. One possibility for improving the hierarchical system is to introduce certain phrase/gap constraints using POS tags or deeper syntax knowledge sources. The long-range POS-based reorderings of verbs used to improve the standard phrase-based system reduced the number of reordering errors, and also the number of lexical errors. A discrepancy considering the missing words can be observed for this method: The human error analysis reports reduction of missing words whereas this is not captured by automatic tools. Deeper analysis regarding different POS classes could possibly reveal more details about this. It also can be seen that the hierarchical

Table 14
Results (raw error counts) of human (left) and automatic (right) error analysis for three different German-to-English WMT systems; Spearman (left) and Pearson (right) correlation coefficients for each translation output across error categories (last column) and for each error category across different translation outputs (last row).

system	infl	order	miss	extra	lex	ρ/r
phrase-based	12 / 32	60 / 235	204 / 199	52 / 40	189 / 521	0.70 / 0.72
+reorder	16 / 44	41 / 212	172 / 200	30 / 56	163 / 495	0.7 / 0.74
hierarchical	17 / 46	100 / 274	107 / 153	68 / 99	171 / 508	0.90 / 0.91
ρ/r	1.00 / 0.90	1.00 / 0.99	0.60 / 0.90	0.5 / 0.62	1.00 / 0.96	

Table 15
Examples of errors in three different German-to-English translation systems: phrase-based, phrase-based with POS-based reorderings, and hierarchical. **Bold words** denote differences between systems.

reference:	The total amount designated for assistance to the system is to be divided into two parts.
phrase-based:	The _____ _____ ________ _____ aid to the system __ __ certain amount in two parts.
+reorder:	_____ To help the system certain total amount __ to be divided into two parts.
hierarchical:	The for the system _____ to help certain total amount will be divided into two parts.
reference:	Retailers are to decide for themselves if they want to pass on the price increases to their customers.
phrase-based:	The retailers themselves must decide whether ___ ___ the _____ doubling their customers **pass**.
+reorder:	Retailers should decide for themselves whether they want **to pass on** increasing costs to their customers.
hierarchical:	The retailers themselves must decide whether they wish **to pass on** to their customers **the price increase**.
reference:	We have made great progress towards an agreement that will be effective on the market, declared the representative of the Bush administration, Henry Paulson.
phrase-based:	We have ____ great progress towards an agreement, the ___ marktwirksam **be**, said the representatives of the administration, **Bush** Henry Paulson.
+reorder:	We have made great progress towards an agreement, which will **be** marktwirksam, said the representatives of the administration, **Bush** Henry Paulson.
hierarchical:	We have made great strides in the direction of an agreement, which will **be** marktwirksam, said the representatives of the **Bush** administration, Henry Paulson.
reference:	The search for Fosset was called off a month after he had disappeared.
phrase-based:	The search for Fosset was a month after whose disappearance **has been abandoned**.
+reorder:	The search for Fosset was **abandoned** a month after his disappearance.
hierarchical:	The search Fosset was a month after the disappearance **have been abandoned**.

system produces fewer lexical errors than the baseline phrase-based system, but more lexical errors than the phrase-based system with reorderings. Syntactic constraints for the hierarchical model might improve this error category as well.

Table 15 presents examples of differences between translation systems and effects of improvements along with some differences between human and automatic error analysis.[4] The first sentence represents an example for the missing words error problem—one can see that there are a number of missing words in the output of the phrase-based system. All of these missing words are detected both by human and

4 It should be noted that it was impossible to show visually all details worthy of explanation.

by automatic error analysis, except for the word *amount*, which is considered by the automatic tools as a reordering error (for the same reason as the example in Table 7). When the reordering technique is applied, the number of errors is reduced and the translation becomes understandable although still not grammatically correct: The sequence *total amount* becomes a reordering error both for humans and for automatic tools. For the words *designed* and *is*, the confusion between missing words (human) and lexical errors (automatic) is present. The hierarchical system generates even fewer missing words, although neither this system nor the phrase-based system with reorderings is able to overcome the reordering problem. It should also be noted that for all three systems, the word *assistance* is detected as a word translated by incorrect lexical choice *aid/to help*, whereas by the humans it is not considered as an error.

The second example illustrates improvements of the standard phrase-based system yielded by POS-based reorderings, as well as advantages and disadvantages of the hierarchical system. The baseline phrase-based system produces several missing words (*they want*, *price increases*), as well as reordering errors (*pass on*). When the reorderings are applied, these errors are no longer present and the overall number of errors is reduced, although there are still some words considered as errors only by automatic tools (i.e., *if* and *price* as lexical errors, *to* as a reordering error). The output of the hierarchical system also does not contain missing words, and the problem of the verb reordering is solved as well (*pass on*), but some other reordering errors are introduced: *the price increase*. It should be noted that for this sentence automatic tools detect that *want* is a lexical error because it is translated as *wish*, whereas the humans of course do not consider this an error.

The third sentence shows advantages of the hierarchical system. The output of the phrase-based system has several missing words (*made*, *that will*) which are not present when the reorderings are applied, and are also not present in the output of the hierarchical system. The same happens with the verb reordering error *be*. However, the noun reordering error *Bush* is not resolved by the POS-based reorderings of verbs, whereas it is at the correct position in the hierarchical system output. Apart from this, all outputs contain an inflectional error *representative/representatives*, as well as a lexical choice error caused by an out-of-vocabulary word *effective on market/marktwirksam*. For the phrase-based output, this error is classified as a lexical error both by humans and by automatic tools, but for the other two outputs, the automatic tools classified the words *effective on* as missing words.

The fourth example shows a case when the automatic error analysis cannot detect differences between translation outputs. The baseline phrase-based output contains verb reordering errors *has been abandoned*. The same error is present in the output of the hierarchical system, whereas the sentence obtained by the phrase-based system with reorderings is completely correct. Nevertheless, these reordering errors are not at all detected by automatic error analysis—it detects reordering errors only if exactly the same words in the reference and in the hypothesis are present. Therefore there are more discrepancies between the human and automatic error analysis for all three outputs: *called off* is detected as missing, *disappeared* as a lexical error, and in the hierarchical system output *have been* are considered as extra words. This example, together with the fact that in general the automatic error analysis detects much higher numbers of lexical errors than the human evaluation, indicates that the automatic error analysis could be improved by using a list of synonyms.

4.3.3 Correlations of the POS Classes Across the Translation Outputs. Correlation coefficients of the POS classes across different TC-STAR and WMT09 translation outputs for each

error category are presented in Table 16. For the cases when both error analyses detected zero errors, the correlation coefficients are omitted. It can be seen that the verbs, nouns, and adverbs have high correlations for each of the error categories (except extra words), as well as that the inflectional and reordering errors have high correlations for almost all POS classes. These error categories and POS classes are used in the further experiments described in the following section.

On the other hand, it can be seen that for the missing words, extra words, and lexical errors, the correlations for some of the POS classes are low or the behavior is dependent on the data set. As already mentioned in Section 4.2, the main source of discrepancies between human and automatic error analysis in general is the difficulty of disambiguation between missing words, lexical errors, and extra words. Nevertheless, a deeper analysis (an "error analysis of error analysis") should be carried out in order

Table 16
Spearman (above) and Pearson (below) correlation coefficients for each POS class within one error category across six different TC-STAR and three different WMT09 English translation outputs. The coefficients are omitted for the cases when both error analyses detected zero errors.

(a) TC-STAR translation outputs

	V	N	A	ADV	PRON	DET	PREP	CON	NUM	PUN
infl	0.96	0.61		1.00						
	0.96	0.02		1.00						
order	0.69	0.76	0.86	0.44	0.54	0.50	0.94	0.66	0.66	0.53
	0.85	0.89	0.98	0.37	0.02	0.66	0.92	0.03	0.03	0.01
miss	0.71	0.76	0.04	0.54	1.00	0.64	0.17	0.87		0.66
	0.67	0.83	−0.15	0.94	0.99	0.61	−0.31	0.88		0.03
ext	0.61	−0.40	0.53	0.86	0.57	0.39	0.21	0.66	0.66	
	0.78	−0.47	0.47	0.95	0.58	0.20	0.01	0.01	0.03	
lex	0.99	1.00	0.83	0.93	0.59	0.47	0.99	0.76		0.66
	0.99	0.99	0.96	0.93	0.98	0.97	0.94	0.71		0.03

(b) WMT09 translation outputs

	V	N	A	ADV	PRON	DET	PREP	CON	NUM	PUN
infl	1.00	0.50								
	0.99	0.01								
order	1.00	1.00	0.88	0.50	0.88	1.00	1.00	0.12	0.88	−0.62
	0.99	0.98	0.69	0.72	0.72	1.00	0.99	−0.28	0.94	−0.92
miss	1.00	0.50	0.12	1.00	1.00	0.50	0.50	−0.12	−0.12	−0.62
	0.99	0.73	0.01	0.97	0.89	0.60	0.95	−0.50	−0.50	−0.94
ext	0.12	0.50	0.50	−0.50	−0.50	0.50	0.50	0.88		0.50
	0.19	0.52	0.01	−0.50	−0.42	0.89	0.94	0.94		−0.06
lex	1.00	0.50	−0.50	1.00	0.12	−1.00	1.00	−1.00	0.62	−0.62
	0.99	0.03	−0.87	1.00	−0.19	−0.98	0.99	−1.00	0.04	−0.66

to better understand all the details. Such an analysis is an interesting and important direction for future work.

5. Comparison of Translation Systems

The experiments in the previous sections showed that the proposed error categories (with the exception of extra words) can be useful for obtaining more information about differences between translation outputs. In this section, we will show some applications. In order to be able to compare translation outputs generated under different conditions (different target languages, different test set sizes, etc.), we introduce word error rates for each error category, that is, we normalize the number of errors over the total reference length. We do not omit the extra words because this category is informative for the distribution of errors in a particular translation output. Thus the following novel metrics are defined:

INFER (**inf**lectional **e**rror **r**ate):
Number of RPER (reference PER) or HPER (hypothesis PER) errors caused by wrong choice of the full word form normalized over the (closest) reference length.

RER (**r**eordering **e**rror **r**ate):
Number of WER errors that do not occur either as RPER (reference PER) or as HPER (hypothesis PER) errors normalized over the (closest) reference length.

MISER (**mis**sing word **e**rror **r**ate):
Number of WER deletions that are not caused by wrong full form choice normalized over the (closest) reference length.

EXTER (**ext**ra word **e**rror **r**ate):
Number of WER insertions that are not caused by wrong full form choice normalized over the (closest) reference length.

LEXER (**lex**ical **e**rror **r**ate):
Number of errors caused neither by wrong full form choice nor by deleting or inserting words normalized over the (closest) reference length.

ΣER (**sum** of **e**rror **r**ates):
Sum of all error categories.

An overview about how these metrics behave in comparison with the standard word error rates WER, PER, and TER along with Spearman and Pearson correlation coefficients is presented in Table 17. The BLEU score is also shown as illustration. All error rates are calculated on the translation outputs analyzed in Section 4. We can see that the sum of all error categories ΣER is always greater than PER, lower than WER, and similar, although in a majority of cases lower, than TER.

In the following experiments, we use the error rates to get more details about differences between

- different language pairs: The error rates are calculated for six translation outputs generated in the framework of the WMT09 shared task in order to see the differences between the target languages as well as the differences in English outputs generated by distinct source languages.

- methods for improvement: Following the ideas from Section 4.3, more details about the effects of POS-based reordering methods are analyzed: local reorderings on the TC-STAR data and long range reorderings on the WMT09 data.

- different translation systems: Outputs generated in the second TC-STAR evaluation by five distinct translation systems (three statistic and two rule-based) are analyzed. In addition, an interesting example from the WMT09 shared task is presented.

Table 17
Error categories — novel error rates (raw error counts in five error categories normalized over reference length) compared with standard word error rates WER, PER, and TER. The BLEU score is also shown as illustration.

(a) Standard error rates (%) (including the BLEU score) and novel error rates.

%	BLEU	WER	PER	TER	INFER	RER	MISER	EXTER	LEXER	ΣER
ArEn (BN)	59.7	29.6	22.4	28.0	1.5	4.2	4.0	8.7	9.3	27.6
ArEn (NW)	72.1	18.8	14.5	17.8	1.1	1.9	4.9	3.6	6.2	17.8
CnEn (NW)	58.0	34.6	21.4	30.0	1.6	6.8	9.6	4.6	9.5	32.1
EsEn 1.3M	60.9	31.1	25.2	29.6	1.0	4.1	4.1	3.2	15.9	28.2
EsEn +reord	61.7	30.0	24.9	28.8	1.0	3.9	4.2	2.7	15.9	27.9
EsEn 13k	48.3	37.6	28.9	36.0	1.0	6.4	4.9	2.9	19.1	34.3
EsEn +reord	51.3	35.8	28.7	34.8	1.1	4.3	5.1	3.4	18.6	32.6
EsEn dict	21.6	57.9	47.7	56.3	1.5	7.5	7.2	2.6	35.4	54.2
EsEn +reord	25.7	54.8	46.8	53.6	1.6	5.5	6.8	2.7	33.4	50.1
DeEn phrase	16.9	66.4	47.5	60.5	2.1	14.8	12.5	2.5	32.8	64.8
DeEn +reord	18.4	65.5	47.2	59.6	2.8	13.3	12.6	3.5	31.2	63.4
DeEn hier	17.2	71.9	48.3	64.7	2.9	17.2	9.6	6.2	32.0	68.0

(b) Spearman (above) and Pearson (below) correlation coefficients between standard error rates (including 1-BLEU) and novel error rates (error categories).

	1-BLEU	WER	PER	TER	INFER	RER	MISER	EXTER	LEXER
ΣER	0.972	1.000	0.937	1.000	0.701	0.930	0.804	−0.302	0.865
	0.982	0.998	0.959	0.990	0.853	0.915	0.814	−0.103	0.914
LEXER	0.834	0.865	0.921	0.865	0.451	0.698	0.582	−0.617	
	0.960	0.926	0.988	0.957	0.604	0.695	0.592	−0.348	
EXTER	−0.274	−0.302	−0.397	−0.302	0.222	−0.089	−0.138		
	−0.172	−0.107	−0.225	−0.146	0.264	0.070	−0.115		
MISER	0.787	0.804	0.598	0.804	0.813	0.843			
	0.749	0.791	0.668	0.744	0.836	0.849			
RER	0.944	0.930	0.804	0.930	0.792				
	0.832	0.902	0.769	0.858	0.905				
INFER	0.726	0.701	0.540	0.701					
	0.771	0.839	0.710	0.794					
TER	0.972	1.000	0.937						
	0.995	0.995	0.987						
PER	0.902	0.937							
	0.988	0.967							
WER	0.972								
	0.986								

5.1 Different Language Pairs

Table 18 presents the error rates and the BLEU scores for six translation outputs from the News domain generated by phrase-based systems in the framework of the WMT09 shared task. The test data used in the WMT shared tasks are very suitable for such comparison because they are parallel for all language pairs. For each language pair only a baseline phrase-based system is used in order to focus only on the language-dependent differences. It can be seen that for all English outputs, independently of the source language, the inflectional error rate (INFER) of about 2% is the smallest error rate. For the other target languages this is not the case: The French and Spanish outputs have an INFER between 6% and 7%, and the German output more than 8%. These results could be expected knowing that the English language is not morphologically rich, whereas Spanish, French, and especially German are.

The highest reordering error rate (RER) is present when translating from and into German, and the highest missing word error rate (MISER) can be observed for the English output generated from the German source text. The category of extra words has been shown not to be reliable for comparison of translation outputs, although it can be seen that all EXTER scores are rather low in comparison to the other error categories. The highest lexical error rate (LEXER) of 35% can be observed for the German output, followed by the English output generated from the German source text (33%). The lexical error rates for the translation from and into Spanish and French are lower and similar, between 29% and 30%.

From the BLEU score and other standard error measures it can be seen that the translation from and particularly into German is the hardest. It can also be observed that translation into French and Spanish is more difficult than translation from these languages into English. The results of the error analysis give more details, for example, such that for translation from and into German language the number of reordering errors is higher than for the other language pairs. Furthermore, when translating from German into English, a high number of missing words should be expected. For translation into German, morphology is an important issue. A high number of inflectional errors is also present for the French and Spanish outputs—higher INFER is the main reason why translation into French and Spanish is more difficult than the other translation direction.

5.2 Methods for Improvement—More Details About POS-Based Reorderings

In order to better understand the reordering problems and improvements obtained by POS-based reordering techniques, we investigate some more details. For the full

Table 18
Error categories for different source and target languages: examples of German-to-English, French-to-English, Spanish-to-English, English-to-German, English-to-French, and English-to-Spanish WMT09 translation outputs.

%	BLEU	INFER	RER	MISER	EXTER	LEXER	ΣER
de-en	17.0	2.0	12.3	13.2	4.4	33.0	65.0
fr-en	23.2	2.2	11.3	9.6	6.1	29.9	59.1
es-en	23.0	2.2	11.3	9.8	5.8	29.6	58.9
en-de	12.7	8.4	12.2	9.9	4.8	35.0	70.3
en-fr	21.7	6.2	11.0	9.7	4.7	29.9	61.4
en-es	21.4	6.9	11.6	8.4	5.1	29.3	61.4

TC-STAR training corpus, we separate the test corpus into two sets—one containing sentences whose source sentences have been actually reordered and the other containing the rest of the sentences. Then we calculate the overall RER measure as well as the RER of noun–adjective groups and the RER of verbs for each of the sets translated by both systems (without and with reorderings). The results in Table 19 show that the overall RER of the reordered set is decreased by the local reorderings whereas for the rest of the sentences a small increase can be observed. Furthermore, it can be noted that for the reordered set the RER of verbs is significantly smaller than the RER of nouns and adjectives which has been improved by local reorderings. For the rest of the sentences there are no significant differences either between RERs of different POS groups or between the system with reorderings and the baseline system. The same tendencies occur for the other translation direction.

In order to better understand the long-range differences in word order, we carried out a similar experiment on the WMT09 German-to-English translation output: We separate the test corpus into two sets, and calculate the RERs. In addition, we also calculate the MISERs, because the category of missing words has also been shown to be rather problematic for the translation from German into English, and to be improved by long-range reorderings. The results are presented in Table 20. It can be observed that the reordered part of the test corpus has a higher RER, which is improved by long-range reorderings. However, the RER of the other part is also indirectly improved. Looking into the specific POS classes, it can be seen that the reordering error rate of nouns and adjectives RER (N,A) is only slightly higher for the reordered sentences than for the rest, whereas the RER (V) is significantly higher for the reordered sentences. When the long-range reorderings are applied, the RER (V) is reduced, but indirectly also the RER (N,A). As for missing words, the overall MISER is similar for both test sets. For the reordered set it is reduced by applying long-range reorderings, and for the rest of the sentences it is slightly increased. The number of missing verbs is much higher for the reordered

Table 19
Effects of local POS-based reorderings on reordering error rates for the Spanish–English translation for reordered sentences and for the rest: overall RER, RER of nouns and adjectives, and RER of verbs.

(a) English output

Spanish→English		RER	RER (N,A)	RER (V)
reordered	baseline	6.0	2.6	0.9
	reorder adjectives	5.4	2.1	0.9
not reordered	baseline	4.2	1.3	1.0
	reorder adjectives	4.3	1.3	1.0

(b) Spanish output

English→Spanish		RER	RER (N,A)	RER (V)
reordered	baseline	6.5	2.3	0.5
	reorder adjectives	6.3	2.2	0.5
not reordered	baseline	4.6	1.2	0.6
	reorder adjectives	4.7	1.2	0.6

Table 20
Effects of long range POS-based reorderings on reordering error rates and missing word error rates for the German–English translation for reordered sentences and for the rest.

(a) Reordering error rates: overall RER, RER of nouns and adjectives, and RER of verbs.

German→English		RER	RER (N,A)	RER (V)
reordered	baseline	12.9	1.9	1.8
	reorder verbs	12.4	1.8	1.6
not reordered	baseline	9.8	1.4	0.7
	reorder verbs	9.5	1.3	0.7

(b) Missing word error rates: overall MISER, MISER of nouns and adjectives, and MISER of verbs.

German→English		MISER	MISER (N,A)/MISER (N)	MISER (V)
reordered	baseline	13.1	2.0 / 2.8	3.5
	reorder verbs	12.4	2.0 / 2.8	2.9
not reordered	baseline	13.6	2.2 / 3.4	2.3
	reorder verbs	13.9	2.1 / 3.5	2.4

set than for the rest, whereas the number of missing nouns and adjectives is similar for both sets. The MISER (V) of the reordered set is significantly reduced by long-range reorderings, whereas the MISER (N,A) remains the same. For the rest of the sentences, there are basically no differences between different POS classes and systems with and without reorderings.

The experiments have also shown that the local reorderings are more targeted to the specific word classes (i.e., nouns and adjectives), improving mostly these words, whereas long-range reorderings do not affect only the words which are actually reordered (in this case: verbs) but they also introduce some indirect improvements, such as reducing errors of other word classes and reducing the number of missing words. This happens due to better alignment learning and better phrase extraction enabled by applying long-range reorderings.

5.3 Different Translation Systems

For the translation outputs analyzed in the previous section, the same phrase-based translation system is used for all experiments. In order to examine how the new error rates reflect the differences between distinct translation systems, we carried out an error analysis of different translation outputs generated by five distinct translation systems in the second TC-STAR evaluation. A total of nine different systems participated in the evaluation, and we selected five representative systems for our experiments which will be referred to as A, B, C, D, and E. For the English language we used the outputs of four systems A, B, C, and D, and for Spanish additionally the output of a system E. The systems A, B, and C are statistical phrase-based and the systems D and E are rule-based.

In Table 21 the new error rates for all translation outputs are presented along with the BLEU score as the official metric of the evaluation. For translation into English, the systems A, B, and C have very similar BLEU scores as well as all error categories. The

Table 21
Error categories for different Spanish-to-English and English-to-Spanish TC-STAR translation systems.

(a) English outputs

English	BLEU	INFER	RER	MISER	EXTER	LEXER	ΣER
A	53.5	2.5	5.8	4.7	4.1	14.5	31.6
B	53.1	2.3	5.7	4.8	3.5	13.9	30.2
C	52.8	2.1	5.9	5.6	3.2	14.1	30.9
D	45.4	2.6	6.9	3.7	5.2	17.5	35.9

(b) Spanish outputs

Spanish	BLEU	INFER	RER	MISER	EXTER	LEXER	ΣER
A	50.0	4.8	5.6	4.6	4.1	15.2	34.4
B	48.2	4.8	5.8	5.3	3.8	15.1	34.8
C	49.6	4.9	5.6	5.3	3.1	14.7	33.7
D	38.9	**5.5**	**6.7**	4.7	4.4	**19.4**	40.7
E	38.6	5.1	**7.8**	4.6	4.7	**19.0**	41.2

worst ranked system according to the BLEU is system D, and from the error rates it can be seen that the main problem for this system is the incorrect lexical choice. The number of reordering errors is also larger for this system than for the others.[5]

For translation into Spanish, the BLEU scores are similar for the three systems A, B, and C with the two systems D and E having lower scores. The error rates show that the main differences between systems A, B, and C on the one side and systems D and E on the other are incorrect lexical choices. The number of reordering errors is also higher for systems D and E, and system D in addition has a higher INFER than the other systems.

WMT09 systems. Another interesting example can be found for the WMT09 French–English translation task of News data: Among twenty participants, the output generated by the Google system[6] has the best BLEU score as well as the best human sentence ranking score (Callison-Burch et al. 2007), whereas the scores for the translation produced by University of Geneva (Wehrli, Nerima, and Scherrer 2009) are the lowest. The sentence rank of a translation system is defined as the percentage of sentences for which this system is judged by human evaluators to be better or equal than any other system. The results for these two systems along with two additional medium ranked systems, the statistical Limsi system and a rule based rbmt3 system, can be seen in Table 22. The BLEU score and the official human rank score are shown along with the five error categories.

Similarly to the TC-STAR systems, the main problem of the worst-ranked systems is the lexical error rate LEXER. The reordering error rate RER also has the same rank as the official overall scores. Nevertheless, the Geneva translation output significantly outperforms the other systems in terms of the inflectional error rate INFER. Looking

5 The number of extra words too, although in previous experiments this error category is shown not to be reliable for system comparison.

6 http://translate.google.com.

Table 22
Human sentence rank (%), BLEU score, and error categories for four French-to-English WMT09 translation systems: the best ranked Google, the worst ranked Geneva, and two medium ranked systems (Limsi (statistical) and rbmt3 (rule-based)).

FrEn	RANK	BLEU	INFER	RER	MISER	EXTER	LEXER	ΣER
google	76.0	31.0	2.1	10.2	9.3	5.2	26.9	53.8
limsi	65.0	26.0	2.3	11.5	6.4	9.4	30.1	59.8
rbmt3	54.0	20.0	2.4	11.9	5.6	11.0	34.7	65.6
geneva	34.0	14.0	**1.5**	12.2	5.8	13.0	**36.9**	69.6

Table 23
Inflectional error rates of verbs, nouns, and adjectives in English outputs produced by four WMT09 translation systems: the best ranked Google, the worst ranked Geneva, and two medium ranked systems (Limsi (statistical) and rbmt3 (rule-based)).

FrEn		google	limsi	rbmt3	geneva
INFER	V	1.4	1.5	1.5	**0.8**
	N	0.7	0.7	0.6	0.5
	A	0.1	0.1	0.1	0.1

into the details about POS classes in Table 23, it can be seen that the INFER of verbs is the main reason for the low INFER of the Geneva translation.

6. Discussion

This work describes a framework for automatic error analysis of translation output that presents just a first step towards the development of automatic evaluation measures which provide partial and more specific information of certain translation problems. The basic idea is to use the actual erroneous words extracted from the standard word error rates WER and PER in combination with linguistic knowledge in order to obtain more information about the translation errors and to perform further analysis of particular phenomena. The overall goal is to get a better overview of the nature of actual translation errors—to identify strong and weak points of a translation system, to get ideas about possible improvements of the system, to analyze improvements achieved by particular methods, and to better understand the differences between different translation systems.

There are many possible ways to carry out automatic error analysis using the proposed framework. The focus of this work is classifying errors into the following five categories: morphological (inflectional) errors, reordering errors, missing words, extra words, and incorrect lexical choice. In addition, the distribution of these error types over POS classes is investigated. This method can be applied to any language pair; the prerequisite is the availability of a morpho-syntactic analyzer for the target language.

The results of the proposed method are compared with the results obtained by human error analysis. Detailed experiments on different types of corpora and various language pairs are carried out in order to investigate two applications of error analysis: estimating the contribution of each error category within one translation output, and comparing different translation outputs using the introduced error categories. For the distribution of error categories within a translation output, we show that the results of

automatic error analysis correlate very well with the results of human error analysis for all translation outputs. In addition, we show that the differences between results of human and automatic error analysis occur mainly due to the difficulty of disambiguation between missing words, lexical errors, and extra words. As for the comparison of different translation outputs, we show that all error categories except extra words correlate well with the human analysis. We also show that verbs, nouns, and adverbs correlate well in all error categories (except extra words), as well as that inflectional and reordering errors correlate well for almost all POS classes. Nevertheless, for missing words, lexical errors, and extra words, some of the POS classes have low correlations. The main reason for these discrepancies is again the problematic disambiguation between the three error categories. A deeper analysis should be carried out in order to understand all details, such as to examine which words/POS classes are classified in one particular category by humans but in another category automatically. Such an "error analysis of error analysis" is an interesting and important direction for future work.

The work described in this article also opens many other directions for future work. The proposed framework can be extended in various ways such as going beyond the word level, introducing deeper linguistic categories, using other alignments apart from WER and RPER/HPER (such as TER, GIZA, etc.), investigating the contribution of source words if source–target alignment information is available, measuring correlations with the human error analysis carried out without reference translations, measuring inter- and intra-annotator agreement for human error analysis and its effects, and so forth.

7. Conclusions

This work presents a first step towards automatic error analysis of machine translation output. We show that the results obtained by the proposed framework correlate very well with the results of human error analysis, as well as that the main source of discrepancies is disambiguation between missing words, lexical errors, and extra words. The new error rates are then calculated for various translation outputs in order to compare them. Different source and target languages are compared to see particular problems for each language pair and translation direction. An analysis of improvements yielded by POS-based reorderings is carried out as well. Finally, we show how the new measures can show differences between distinct translation systems, namely, what are the weak/strong points of particular systems. The presented framework offers a number of possibilities for future work, both to improve the proposed metrics as well as to investigate other set-ups.

Acknowledgments

This work was partly funded by the European Union under the integrated project TC-STAR — Technology and Corpora for Speech to Speech Translation (IST-2002-FP6-506738); by the Quaero Programme, funded by OSEO, French State agency for innovation; and by the Defense Advanced Research Project Agency (DARPA) under contract No. HR0011-06-C-0023. Any opinions, findings, conclusions, or recommendations expressed in this work are those of the authors and do not necessarily reflect the views of DARPA.

Special thanks to Adrià de Gispert, Deepa Gupta, Patrik Lambert, and Necip Fazil Ayan.

References

Banerjee, Satanjeev and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgements. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, pages 65–72, Ann Arbor, MI.

Callison-Burch, Chris, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (Meta-)Evaluation of machine translation. In *Proceedings of the 2nd ACL 07 Workshop on Statistical Machine Translation (WMT 07)*, pages 136–158, Prague.

Callison-Burch, Chris, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of the 4th EACL 09 Workshop on Statistical Machine Translation (WMT 09)*, pages 1–28, Athens.

Chiang, David. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228.

Doddington, George. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the ARPA Workshop on Human Language Technology*, pages 128–132, San Diego, CA.

Giménez, Jesús and Enrique Amigó. 2006. IQMT: A framework for automatic machine translation evaluation. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 06)*, pages 685–690, Genoa.

Kirchhoff, Katrin, Owen Rambow, Nizar Habash, and Mona Diab. 2007. Semi-automatic error analysis for large-scale statistical machine translation. In *Proceedings of the MT Summit XI*, pages 289–296, Copenhagen.

Leusch, Gregor, Nicola Ueffing, and Hermann Ney. 2006. CDER: Efficient MT evaluation using block movements. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 06)*, pages 241–248, Trento.

Leusch, Gregor, Nicola Ueffing, David Vilar, and Hermann Ney. 2005. Preprocessing and normalization for automatic evaluation of machine translation. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, pages 17–24, Ann Arbor, MI.

Levenshtein, Vladimir Iosifovich. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8):707–710.

Llitjós, Ariadna Font, Jaime G. Carbonell, and Alon Lavie. 2005. A framework for interactive and automatic refinement of transfer-based machine translation. In *Proceedings of the 10th Annual Conference of the European Association for Machine Translation (EAMT 05)*, pages 87–96, Budapest.

Lopez, Adam and Philip Resnik. 2005. Pattern visualization for machine translation output. In *Proceedings of HLT/EMNLP on Interactive Demonstrations*, pages 12–13, Vancouver.

Matusov, Evgeny, Gregor Leusch, Oliver Bender, and Hermann Ney. 2005. Evaluating machine translation output with automatic sentence segmentation. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT 05)*, pages 148–154, Pittsburgh, PA.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wie-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 02)*, pages 311–318, Philadelphia, PA.

Popović, Maja, Adrià de Gispert, Deepa Gupta, Patrik Lambert, Hermann Ney, José B. Mariño, Marcello Federico, and Rafael Banchs. 2006. Morpho-syntactic information for automatic error analysis of statistical machine translation output. In *Proceedings of the 1st NAACL 06 Workshop on Statistical Machine Translation (WMT 06)*, pages 1–6, New York, NY.

Popović, Maja and Hermann Ney. 2006a. POS-based word reorderings for statistical machine translation. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 06)*, pages 1278–1283, Genoa.

Popović, Maja and Hermann Ney. 2006b. Statistical machine translation with a small amount of bilingual training data. In *Proceedings of the LREC 06 Workshop on Strategies for Developing Machine Translation for Minority Languages*, pages 25–29, Genoa.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation error rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA 06)*, pages 223–231, Boston, MA.

Turian, Joseph, Luke Shen, and I. Dan Melamed. 2003. Evaluation of machine translation and its evaluation. In *Proceedings of the MT Summit IX*, pages 23–28, New Orleans, LA.

Vilar, David, Evgeny Matusov, Saša Hasan, Richard Zens, and Hermann Ney. 2005.

Statistical Machine Translation of European Parliamentary Speeches. In *Proceedings of the MT Summit X*, pages 259–266, Phuket.

Vilar, David, Jia Xu, Luis Fernando D'Haro, and Hermann Ney. 2006. Error analysis of statistical machine translation output. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 06)*, pages 697–702, Genoa.

Wehrli, Eric, Luka Nerima, and Yves Scherrer. 2009. Deep linguistic multilingual translation and bilingual dictionaries. In *Proceedings of the 4th EACL 09 Workshop on Statistical Machine Translation (WMT 09)*, pages 90–94, Athens.

Xiong, Deyi, Min Zhang, AiTi Aw, and Haizhou Li. 2010. Linguistically annotated reordering: Evaluation and analysis. *Computational Linguistics*, 36(3):535–568.

Zens, Richard, Franz Josef Och, and Hermann Ney. 2002. Phrase-based statistical machine translation. In *25th German Conference on Artificial Intelligence (KI2002)*, pages 18–32, Aachen.

Zhou, Ming, Bo Wang, Shujie Liu, Mu Li, Dongdong Zhang, and Tiejun Zhao. 2008. Diagnostic evaluation of machine translation systems using automatically constructed linguistic check-points. In *Proceedings of the 22nd International Conference on Computational Linguistics (CoLing 2008)*, pages 1121–1128, Manchester.